HELLENIC REPUBLIC
National and Kapodistrian
University of Athens
EST. 1837

National and Kapodistrian University of Athens

School of Health Sciences

Department of Pharmacy

Laboratory of Biopharmaceutics and Pharmacokinetics

Machine Learning Methods in Bioequivalence and Clinical Studies

Doctoral Thesis

Papadopoulos Dimitrios

Athens, 2024

From error to error, one discovers the entire truth.

Sigmund Freud

To my family, Evi, Giorgos and Athina who have supported and continue to support me throughout this journey called “life”.

The approval of this PhD dissertation by the Department of Pharmacy University of the National and Kapodistrian University of Athens does not imply acceptance of the author's opinions. (L. 5343/1932, article 202).

THESIS EVALUATION BOARD MEMBERS

- **Vangelis Karalis** (Academic Supervisor)

Associate Professor, Department of Pharmacy, National and Kapodistrian University of Athens, Greece

- **Georgia Karali** (Member of the Advisory committee)

Associate Professor, Department of Mathematics, National and Kapodistrian University of Athens, Greece

- **Evangelos Terpos** (Member of the Advisory committee)

Professor, School of Medicine, National and Kapodistrian University of Athens, Greece

- **Aleksandra Catic-Djordjević**

Associate Professor, Department of Pharmacy, University of Nis, Serbia

- **Ioannis Dotsikas**

Associate Professor, Department of Pharmacy, National and Kapodistrian University of Athens, Greece

- **Sophia Markantonis-Kyroudi**

Emeritus Professor, Department of Pharmacy, National and Kapodistrian University of Athens, Greece

- **Anastasia Pippa**

Assistant Professor, Department of Pharmacy, National and Kapodistrian University of Athens, Greece

Acknowledgements

Completing this PhD has been a challenging yet an incredibly rewarding journey, and I am deeply grateful to all those who have supported me along the way.

First and foremost, I would like to express my sincere gratitude to my supervisor, Associate Professor Vangelis D. Karalis, for his unwavering support, guidance and encouragement throughout my research. He is undoubtfully one of the best mentors I ever had, a patient tutor, an inspiring professor, a knowledgeable scientist and a kindhearted person. His expertise and feedback have been invaluable in shaping this dissertation. Similarly, I would like to express my appreciation to Associate Professor Georgia Karali, member of my academic advisory committee, whom I was fortunate to meet in my first ever university lecture during my bachelor studies and with whom we collaborated and wrote an excellent publication. I would further like to express my gratitude to Professor Evangelos Terpos, also member of my academic advisory committee, who very much assisted me and provided the data for my first publication.

I would also like to express my sincere and heartfelt gratitude to Associate Professor Aleksandra Catic-Djordjević, Associate Professor Ioannis Dotsikas, Emeritus Professor Sophia Markantonis-Kyroudi and Assistant Professor Anastasia Pippa for the time and effort they dedicated to providing feedback and suggestions for improvements to my dissertation, and for honoring me by serving as members of my PhD committee.

Last but not least, I would like to express my deep gratitude and love, to my family, for their constant encouragement and support, without whom this journey would not have been possible.

Table of contents

Περίληψη

Εισαγωγή

Ο στόχος αυτής της διατριβής ήταν να εισαγάγει την ιδέα της εφαρμογής μεθόδων μηχανικής εκμάθησης και νευρωνικών δικτύων, των πιο προηγμένων μοντέλων τεχνητής νοημοσύνης για τη βελτίωση των διαδικασιών και των αποτελεσμάτων των κλινικών δοκιμών καθώς και των μελετών βιοϊσοδυναμίας. Ο σκοπός ήταν να χρησιμοποιηθούν αυτές οι υπολογιστικές τεχνικές για να να βελτιωθεί η ακρίβεια και η ισχύς των αναλύσεων σε αυτούς τους σημαντικούς τομείς της ιατρικής έρευνας. Η διατριβή αυτή ξεκίνησε την περίοδο της πανδημίας του κορονοϊού και στα πλαίσια αυτά οι πρώτες δύο ερευνητικές εργασίες αφορούσαν τη διερεύνηση της κινητικής των εξουδετερωτικών αντισωμάτων (Nabs) χρησιμοποιώντας ένα κινητικό μοντέλο, και στη διερεύνηση των ατομικών χαρακτηριστικών που θα μπορούσαν να προβλέψουν τα επίπεδα των NAbs εφαρμόζοντας τέσσερις μεθόδους μηχανικής εκμάθησης. Στις επόμενες τρεις ερευνητικές εργασίες αναπτύχθηκε μία μέθοδος αύξησης δεδομένων που συνδύαζε παραγωγικά νευρωνικά δίκτυα και προσομοιώσεις Monte Carlo. Η μεθοδολογία αυτή εφαρμόστηκε στο πλαίσιο κλινικών μελετών και μελετών βιοϊσοδυναμίας για φάρμακα χαμηλής, μεσαίας και υψηλής μεταβλητότητας.

Μέθοδοι

Αναφορικά με τη μελέτη της κινητικής των Nabs, αναπτύχθηκε ένα μοντέλο που περιλάμβανε ένα διαμέρισμα (ολόκληρο το σώμα) για να περιγράψει τα επίπεδα των αντισωμάτων μετά τον εμβολιασμό, ενώ παράλληλα εφαρμόστηκαν τέσσερις τεχνικές μηχανικής μάθησης. Συγκεκριμένα, ανάλυση κύριων συνιστωσών και ανάλυση παραγόντων μικτών επιδράσεων για τις αλληλεπιδράσεις μεταξύ των διαφορετικών χαρακτηριστικών και πώς αυτά επηρεάζουν τα επίπεδα των NAbs, και K-means και random forest (τυχαίο δάσος) για να ομαδοποιηθούν τα άτομα σε διακριτές ομάδες και να ποσοτικοποιηθεί η επίδρασή τους. Όσον αφορά την εισαγωγή της ιδέας εφαρμογής παραγωγικών νευρωνικών δικτύων για την αναγέννηση εικονικών εθελοντών και τη μείωση του απαιτούμενου μεγέθους δείγματος στις κλινικές μελέτες και τις μελέτες βιοϊσοδυναμίας, αναπτύχθηκε μια μεθοδολογία που συνδύαζε προσομοιώσεις Monte Carlo και variational autoencoders (VAEs). Οι προσομοιώσεις Monte Carlo χρησιμοποιήθηκαν για να μιμηθούν τις ακριβείς συνθήκες της κλινικής μελέτης ενώ το VAE εφαρμόστηκε σε ένα υποσύνολο του αρχικού δείγματος, για να δημιουργήσει νέα, συνθετικά δεδομένα, βασισμένα στα πραγματικά δεδομένα, σε μία σειρά από σενάρια και διαφορετικές συνθήκες.

Αποτελέσματα

Στην περίπτωση της μελέτης της κινητικής των εξουδετερωτικών αντισωμάτων, το αναπτυχθέν μοντέλο αποκάλυψε ότι υπάρχουν τρεις διακριτές κινητικές φάσεις για τον χρόνο που έχει περάσει από τον εμβολιασμό και ότι τα NAbs εξαφανίζονται σχετικά αργά στην αρχή, αλλά η απομάκρυνσή τους γίνεται περίπου 6 φορές ταχύτερα από τον τρίτο έως τον έκτο μήνα. Εντοπίστηκαν πέντε ομάδες ατόμων με διακριτά χαρακτηριστικά ενώ χρησιμοποιώντας δύο κύριες συνιστώσες, εξηγήθηκε το 63.4% της μεταβλητότητας και εντοπίστηκε η θετική συσχέτιση μεταξύ των επιπέδων των NAbs στους 3 μήνες (M3) και στους 9 μήνες (M9) μετά τον εμβολιασμό, το οποίο επαληθεύεται από το random forest καθώς το M3 φαίνεται να είναι ο πιο σημαντικός παράγοντας στην πρόβλεψη του επιπέδου των NAbs 9 μήνες μετά τον εμβολιασμό.

Όσον αφορά την εφαρμογή παραγωγικών νευρωνικών δικτύων στις κλινικές μελέτες με σκοπό τη μείωση του απαιτούμενου μεγέθους δείγματος, αναπτύσσοντας μία πρωτότυπη μεθοδολογία και εξετάζοντας διάφορες μορφές του παραγωγικού νευρωνικού δικτύου, VAE, δημιουργήθηκαν δείγματα με εικονικούς εθελοντές παρόμοια με τα πραγματικά. Τα συνθετικά αυτά δεδομένα απέδωσαν εξίσου καλά, αν όχι καλύτερα, με τα πραγματικά δεδομένα ακόμη και στην περίπτωση που χρησιμοποιήθηκε μόνο το 30-40% των πραγματικών δεδομένων. Αξίζει να σημειωθεί ότι σε σενάρια με υψηλή μεταβλητότητα, τα δεδομένα που δημιουργήθηκαν από το VAE παρουσίασαν υψηλότερη στατιστική ισχύ, μειώνοντας αποτελεσματικά τον "θόρυβο", ενισχύοντας την αξιοπιστία των αποτελεσμάτων.

Εξίσου καλά ήταν τα ευρήματα και στις μελέτες βιοϊσοδυναμίας. Εφαρμόζοντας την ίδια μεθοδολογία, και διερευνώντας διαφορετικές παραμέτρους για τα VAEs και εξετάζοντας πολλαπλά σενάρια, συμπεριλαμβανομένων διαφορετικών επιπέδων μεταβλητότητας, αρχικών μεγεθών δείγματος, μεγεθών δείγματος που δημιουργήθηκαν από VAE και διαφορών απόδοσης μεταξύ των φαρμακευτικών προϊόντων που συγκρίνονται, αποδείχθηκε ότι χρησιμοποιώντας παραγωγικούς αλγορίθμους, και πιο συγκεκριμένα VAEs, κατέστη δυνατό να επιτευχθούν τα ίδια και σε πολλές περιπτώσεις καλύτερα αποτελέσματα, με αρκετά μικρότερο δείγμα από το αρχικό, μειώνοντας έτσι σημαντικά το κόστος και τον χρόνο που απαιτείται για την ολοκλήρωση των μελετών. Ειδικότερα στην περίπτωση των φαρμάκων με υψηλή μεταβλητότητα, τα συνθετικά δεδομένα, είχαν τουλάχιστον παρόμοια με τα πραγματικά, χρησιμοποιώντας μικρότερο δείγμα, ακόμη και όταν η σύγκριση γινόταν με τα κλιμακούμενα όρια βιοϊσοδυναμίας που προτείνονται σήμερα. Δηλαδή, η χρήση παραγωγικών νευρωνικών δικτύων δεν έχει ανάγκη την χρήση κλιμακούμενων ορίων ή ειδικών κριτηρίων και έχει καθολική εφαρμογή.

Συζήτηση

Συνολικά, η μοντελοποίηση της κινητικής των NAbs έδειξε ότι υπάρχουν τρεις διακριτές κινητικές φάσεις για τον χρόνο που έχει περάσει από τον εμβολιασμό και ότι τα NAbs εξαφανίζονται σχετικά αργά στην αρχή, αλλά η απομάκρυνσή τους γίνεται περίπου 6 φορές ταχύτερα από τον τρίτο έως τον έκτο μήνα. Η ανάλυση κύριων συνιστωσών έδειξε τη στενή σχέση μεταξύ M3 και M9 και η ανάλυση μικτών δεδομένων αποκάλυψε ότι η παχυσαρκία και η ηλικία έχουν αρνητική επίδραση στα επίπεδα των NAbs, ενώ το φύλο δεν είχε καμία επίδραση σε καμία από τις πέντε διακριτές ομάδες που εντοπίστηκαν από την K-means. Τα επίπεδα των NAbs σε διαφορετικές χρονικές περιόδους μετά τον εμβολιασμό είναι πιο σημαντικά από την ηλικία, το φύλο και τον δείκτη μάζας σώματος, για τα επίπεδα των NAbs μετά από 9 μήνες από τον εμβολιασμό, σύμφωνα με το random forest.

Η εισαγωγή της χρήσης παραγωγικών νευρωνικών δικτύων στις κλινικές μελέτες με σκοπό τη μείωση του μεγέθους δείγματος και συνδυάζοντας προσομοιώσεις Monte Carlo με VAE, αποδείχθηκε ότι τα VAE μπορούν να αποτελέσουν ένα πολύτιμο εργαλείο στις κλινικές δοκιμές και στις μελέτες βιοϊσοδυναμίας. Χρησιμοποιώντας τα VAE, η στατιστική ισχύ μπορεί να αυξηθεί ενώ το απαιτούμενο μέγεθος μιας κλινικής μελέτης, μπορεί να μειωθεί έως και 30%, μειώνοντας έτσι το απαιτούμενο μέγεθος δείγματος και συνεπώς μειώνοντας το κόστος και τον χρόνο, ενώ παράλληλα αντιμετωπίζουν ηθικά ζητήματα που σχετίζονται με τη συμμετοχή ανθρώπων.

Συμπεράσματα

Συνολικά, αυτή η διατριβή έδειξε ότι η μηχανική μάθηση επιτρέπει την αναγνώριση σύνθετων μοτίβων και τάσεων που είναι δύσκολο να ανιχνευθούν με άλλους τρόπους, στον τομέα των κλινικών μελετών. Χρησιμοποιώντας μεθόδους μηχανικής εκμάθησης εντοπίστηκαν αλληλεπιδράσεις μεταξύ των χαρακτηριστικών ενός ατόμου και των επιπέδων των NAbs μετά από 9 μήνες από τον εμβολιασμό και ποσοτικοποιήθηκε η επίδρασή τους.

Το πιο αξιοσημείωτο είναι ότι αυτή η διατριβή προτείνει για πρώτη φορά τη χρήση των VAEs με σκοπό την αύξηση των δεδομένων σε κλινικές και μελέτες βιοϊσοδυναμίας και τη μείωση του απαιτούμενου μεγέθους δείγματος. Έδειξε ότι η εφαρμογή των VAEs σε κλινικές και μελέτες βιοϊσοδυναμίας αντιπροσωπεύει ένα σύγχρονο και χρήσιμο εργαλείο που μπορεί να μειώσει σημαντικά την ανάγκη για μεγάλο αριθμό ανθρώπων, να μειώσει το κόστος και να συντομεύσει τους χρόνους ολοκλήρωσης των κλινικών δοκιμών, διατηρώντας ή ακόμα και βελτιώνοντας την ποιότητα και την αξιοπιστία των αποτελεσμάτων.

Abstract

Introduction

This dissertation aimed to utilize machine learning and neural networks, the forefront models in deep learning, to enhance the processes and outcomes of clinical trials as well as bioequivalence studies. The goal was to use these advanced computational techniques to gain deeper insights and improve the accuracy and efficiency of the analyses in these important areas of medical research.

To achieve that, the first two research papers focused on an analysis of the kinetics of neutralizing antibodies (NAbs) against SARS-CoV-2 using a kinetic model and identifying their predictive factors by utilizing four machine learning algorithms. In the subsequent three research papers, a novel data augmentation framework was developed, which utilizes generative neural networks and Monte Carlo simulations. The framework was applied in the context of clinical trials bioequivalence testing of low, mid and highly variable drugs. The framework aims to reduce the required sample size for this type of studies which brings numerous breakthrough benefits.

Methods

To investigate the kinetics of NAbs, a kinetic model was used to describe their elimination. The optimal model included a single compartment (the whole body) and linear elimination kinetics. To identify the individuals' characteristics that could predict the NAbs levels, four machine learning techniques were applied. Namely principal component analysis and factor analysis of mixed data, K-means clustering and random forest. The first two methods revealed the interactions between different features and how these affect the Nabs levels whereas the last two allowed us to group the individuals into distinct groups and quantify the predictive factors of NAbs respectively.

Concerning the use of generative algorithms in clinical research, a framework was developed that combined Monte Carlo simulations with a generative neural network, namely variational autoencoders. The Monte Carlo simulations were utilized to replicate the exact conditions of the clinical trial and the BE study, whereas the VAE was applied to a subsample of the original dataset, to generate new, synthetic data, based on the real ones. Various scenarios were tested and different hyperparameters of the VAE model were explored to achieve the optimal model.

Results

The kinetic model identified three distinct kinetic phases on the time elapsed since vaccination and that the NAbs disappear relatively slow at first, but that their removal becomes around six times greater from the third to the sixth month, indicating that they are eliminated much more quickly. K-means identified five unique groups of individuals, each one driven by unique characteristics, whereas using two principal components, were able to explain 63.4% of the variability and identify the positive relation between the NAbs levels at 3 months (M3) and 9 months (M9) after vaccination. This was validated by the random forest, which indicated that the NAbs levels after 3 months is the most important feature when predicting the NAbs levels after 9 months.

To reduce the required sample size in clinical studies, an innovative methodology was used and by utilizing various forms of VAEs, we were able to create virtual samples very similar to the real ones. These synthetic data performed at least as well as the real data, even when only 30-40% of the real data was used. It is worth noting that in scenarios with high variability, the data generated by the VAE showed higher statistical power, effectively reducing "noise", and improving the reliability and robustness of the results. The findings in bioequivalence studies were very desired as well. By applying the same methodology, investigating different parameters for VAEs, and testing multiple scenarios, including different levels of variability, original sample sizes, sample sizes generated by the VAE, and average performance differences between the pharmaceutical products being compared, it was demonstrated that using generative algorithms, and more specifically VAEs, we can achieve the same and in many cases better results, with a significantly smaller sample than the original, thus significantly reducing the cost and time required to complete the studies. Particularly in the case of high variability drugs, the synthetic data performed similarly to the real data, using a smaller sample, even without scaling the confidence interval limits.

Discussion

Overall, the modeling of NAb kinetics showed that there are three distinct kinetic phases based on the time since vaccination. Initially, NAbs decline relatively slowly, but their clearance rate increases approximately sixfold from the third to the sixth month. Principal components analysis showed the strong relationship between M3 and M9 and factor analysis of mixed data revealed that obesity and age have a negative effect to the NAbs levels whereas gender did not have any effect in any of the five distinct groups that were identified by K-means. Random forest indicates that the NAbs levels at different time periods after the vaccination, are more important than age, gender and BMI, when predicting the NAbs levels after 9 months of the vaccination.

The optimized VAEs demonstrated superior performance than the subsampled and similar and many times better than the original datasets, indicating that similar BE testing results, can be achieved by using less samples.

The introduction of generative neural networks in clinical studies to reduce sample sizes, combined with Monte Carlo simulations and VAE, demonstrated that VAEs can serve as a valuable tool in clinical trials and bioequivalence studies. Using VAEs, statistical power can be increased while the required size of a clinical study can be reduced by up to 30%, thereby lowering the necessary sample size, reducing costs and time, and addressing ethical issues related to human participation.

Conclusions

Overall, this dissertation demonstrated that machine learning enables the identification of complex patterns and trends that are difficult to detect by other means in clinical studies. Using machine learning methods, interactions were identified between an individual's characteristics and NAb levels nine months after vaccination, and their impact was quantified. Most notably, this dissertation proposes, for the first time, the use of VAEs to augment data in clinical and bioequivalence studies and to reduce the required sample size. It showed that applying VAEs in clinical and bioequivalence studies represents a modern and useful tool that can significantly reduce the need for large sample sizes, lower costs, and shorten the completion times of clinical trials, while maintaining or even enhancing the quality and reliability of results.

Chapter A: Introduction

A1. Artificial Intelligence

Artificial intelligence (AI) refers to the conceptualization and design of computer systems that can handle tasks that normally require human intelligence. AI is a broad term that includes a variety of technologies, such as machine learning, neural networks and deep learning. AI has gained popularity the last couple of years, however the formal study of AI started in 1950, when Alan Turing published a seminar paper titled "Computing Machinery and Intelligence" [1]. That paper laid the groundwork for the field of AI. After some challenges in the 1970s and 1980s, AI research continued to reach its peak in the 21st century, due to development in machine learning, neural networks and computational power, with tons of applications in different fields [2].

In logistics, AI facilitated the automation of well-defined workflows and improve supply chain decision making [3, 4]. In social sciences AI can build on the existing research in psychology and cognitive science, regarding how people operate, and provide some insights [5] and assist on human decision making [6].

The advantages of AI in healthcare and medicine have been excessively discussed and a lot of articles provide overviews of the medial AI research [7-9]. AI can assist in providing personalized medical diagnosis [10], assist radiologist and pathologists [11]. AI enabled a lot of companies in the pharmaceutical sector to accelerate the process of discovering drugs [12] by streamlining the process and reducing repeated work [13] and assists in drug repurposing [14].

Regardless its popularity AI raises ethical considerations, especially in healthcare. In the late 2010s, the funding of companies analyzing images using AI, surpassed $1.2 billion [15], though, there are still some concerns regarding the anonymity of the patients and how this is and the bias of the algorithms [16]. To overcome these obstacles, AI systems need to be transparent, accessible and mitigate and detect biases following the standards of medical ethics [17].

Despite the forementioned attention points, AI is a rapidly evolving field, with the ability to revolutionize and enhance numerous facets of human life. It is crucial for researchers, policy makers and society to cooperate to harness its potential.

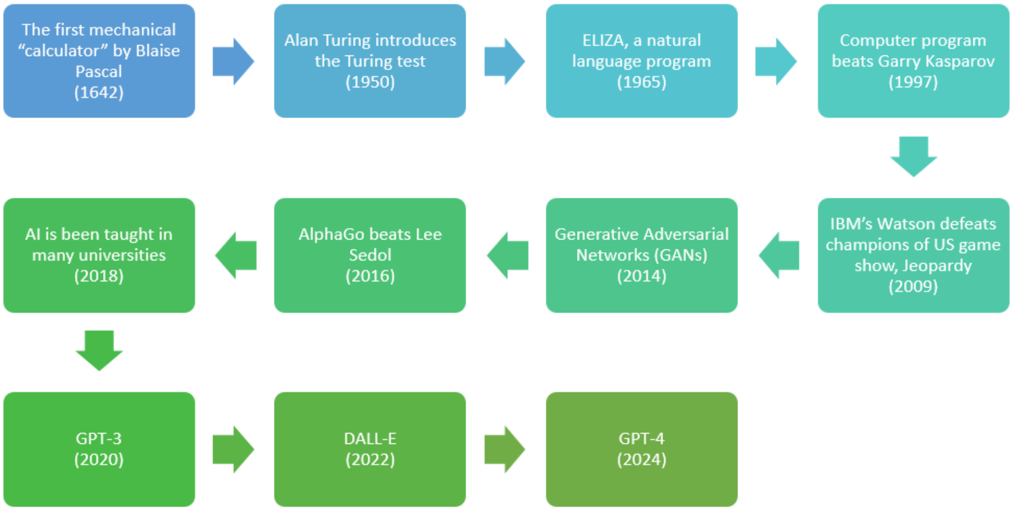

The first mechanical "calculator" by Blaise Pascal (1642)
Alan Turing introduces the Turing test (1950)
ELIZA, a natural language program (1965)
Computer program beats Garry Kasparov (1997)
IBM's Watson defeats champions of US game show, Jeopardy (2009)
Generative Adversarial Networks (GANs) (2014)
AlphaGo beats Lee Sedol (2016)
AI is been taught in many universities (2018)
GPT-3 (2020)
DALL-E (2022)
GPT-4 (2024)

Figure 1. The timeline of Artificial Intelligence. Adapted from [18].

A2. Machine Learning

Machine learning (ML) is a branch of AI. ML is combining statistics and information technology [19] and focuses on the development of algorithms and statistical models. ML models have numerous applications in finance, spacecraft engineering and computer vision, biomedical and healthcare applications [20]. Since data have become easier to get (i.e. big-data era) ML models have gained popularity the last few years.

These models improve automatically through experience, via a process called "learning". During the learning phase, the model takes as input some "data" (i.e. training data) and applies on them a mathematical procedure (i.e. algorithm) to output the "model". The model is a set of mathematical rules and equations that are derived by analyzing the data [21, 22]. Finally, the model's performance is assessed on some new data (i.e. test data) with the use of an appropriate evaluation metric.

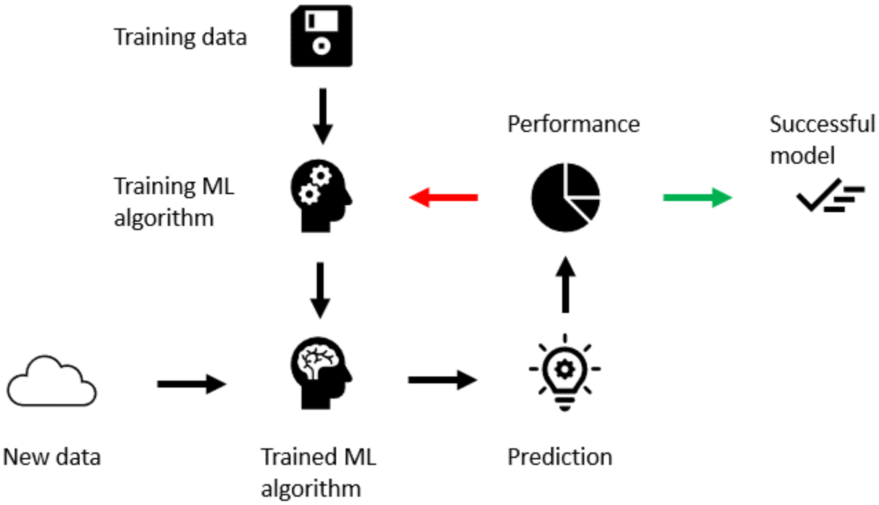


Figure 2. The process of training a Machine Learning model. The model is trained using the training data and its performance is evaluated using new data until its performance is satisfactory. Adapted from [23].

Depending on the application, the data can come from different sources, like databases or the world wide web. Data can be structured (e.g. xlsx files), semi-structured (e.g. XML files) or unstructured (e.g. text and images). Data can be labeled or unlabeled, meaning that either they are tagged with one or more labels to provide context or meaning, or not. Depending on the type of data and the research, a suitable type of machine learning algorithm is applied. Thus, for labeled data a supervised machine learning algorithm is used, for unlabeled data an unsupervised ML algorithm is used.

Supervised learning is the most prevalent form of machine learning. To train a supervised machine learning model a labeled dataset is used, where each input is paired with a corresponding output. This allows the model to predict the output for new inputs accurately. The most common and simplest algorithm used in supervised learning is linear regression. There are more complex supervised ML algorithms, like support vector machines and neural networks.

Supervised machine learning use cases are split into two categories, regression and classification. Both problems involve predicting an output based on the input, but the nature of the output is different. Regression is when the desired output is continuous and numerical (e.g. sales numbers or house pricing) whereas classification is when the output variable is categorical (e.g. “Yes” and “No”, or “Red” and “Green” and “Blue”). When the output variable has two categories, then it is called binary classification and multi-class otherwise. The evaluation metric is different between regression and classification problems. For the first, the mean squared error (MSE) is commonly used whereas for the latter usually accuracy is used.

One key challenge in supervised learning problems is “overfitting”. Overfitting is when the model captures well the patterns in the training dataset, but fails to do the same in test data. There are commonly used techniques to prevent overfitting, such as cross-validation and regularization. During cross-validation, the dataset is split into multiple subsets. The model is trained on each subset to ensure it generalizes well. On the other hand, regularization favors simpler models by penalizing the more complex models, to avoid fitting in the noise in the data.

Unsupervised learning leverages unlabeled data. The model finds patterns and relationships within the data without any guidance on what the output should be. This type of learning is often used for clustering. Clustering algorithms, such as k-means and hierarchical clustering, group similar data points together by identifying relationships between the variables of the dataset. More advanced algorithms include Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Models (GMM).

Another challenge that both supervised and unsupervised applications have, is the features selection, that is the choice of appropriate features to be included model. Before using the appropriate, usually new features need to be created, modified and selected from the raw data. This procedure is called “feature engineering” and is an important step of machine learning which significantly improves the model’s performance. Techniques such as normalization, standardization, and dimensionality reduction are usually included in that step. Normalization scales the features to a specific range, while standardization transforms them to follow a standard Gaussian distribution. Dimensionality reduction methods’ goal is to reduce the number of features while preserving as much information as possible. The most common are principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) [24].

In semi-supervised machine learning labeled data are combined with unlabeled. Usually, the size of labeled data is smaller. This approach is extremely useful when labeling data is either labor-intensive or costly [25]. Image recognition, natural language processing, and are applications where semi-supervised is used.

Lastly, reinforcement learning is a type of machine learning where an algorithm learns to make decisions by experimenting and learning from its mistakes. The computer is either rewarded or penalized through feedback, which leverages for its future decisions. This approach is particularly effective in scenarios where the optimal solution is not known and must be discovered through interaction, such as in game playing, robotics, and autonomous driving [26].

Machine learning is applied in a wide range of industries. In healthcare, there are numerous applications, like diagnosing diseases and creating treatment plans that are personalized. For example, brain tumors can be classified by analyzing medical images and leveraging machine learning models [27] or in general analyze medical images [28]. In finance, machine learning algorithms are employed for fraud detection [29, 30], algorithmic trading [31], and credit scoring [32]. These models can identify suspicious transactions, optimize trading strategies, and assess the creditworthiness of individuals and businesses.

In the entertainment industry, companies like Netflix, Spotify and YouTube, use machine learning to create recommendation systems that suggest movies, music, and other content based on user preferences [33, 34]. In the car industry, Tesla leverages machine learning to build self-driving cars, able to understand the environment they are into, empower them to decide, and navigate safely [35].

Machine learning is widely used, and despite its many successes, it should take into consideration ethical, safety and social concerns. One major issue is bias in machine learning models. The model will be unable to generalize properly in case the dataset is biased, which will lead into undesirable results. For example, facial recognition algorithms tend have higher error rates for certain demographic groups [36, 37], in 2016 a “designer error” in Autopilot may have led to a tragic incident [38]. To account for this, researchers and machine learning practitioners are developing methods to detect and mitigate bias in machine learning models, such as fairness-aware algorithms and diverse training datasets.

Another concern is the interpretability of machine learning models. Many critical applications, especially in healthcare and finance require an understanding of the rationale of a decision before this is made. Thus, the need for explainable machine models, becomes imperative. A lot of efforts have been made on this topic [39, 40] to promote machine learning models that are transparent with respect to their decision-making mechanism [41].

In summary, the field of machine learning is undergoing rapid evolution, significantly impacting various industries and demonstrating substantial potential for future advancements. Leveraging statistical algorithms based on data and continuously enhance their performance new opportunities for automation and decision making will arise. Nevertheless, it also brings forth challenges and ethical concerns that require careful consideration to ensure its fair and equitable application. As ongoing research and development continue to progress, machine learning is going to be a cornerstone in determining the future.

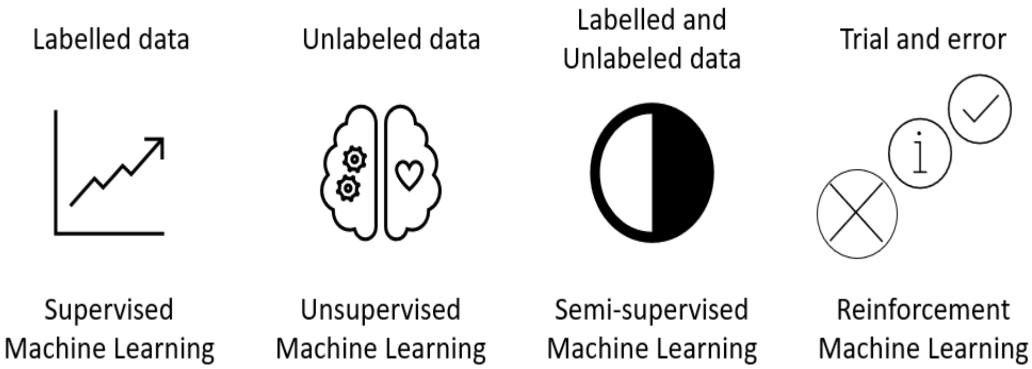

Labelled data
Unlabeled data
Labelled and Unlabeled data
Trial and error
Supervised Machine Learning
Unsupervised Machine Learning
Semi-supervised Machine Learning
Reinforcement Machine Learning

Figure 3. Types of Machine Learning. Each type of Machine Learning model is used for different types of data.

A3. Deep learning and neural networks

Neural networks (NNs), a branch of machine learning, form the core of deep learning algorithms. Their design and function are modeled after the human brain, imitating the way biological neurons communicate through neural connections [42, 43].

The structure of the NN can highly differ from one application to another. Some neural networks are more complex (i.e. deep) and have more connections (i.e. fully connected) than others (i.e. sparsely connected). NNs are used in both supervised (regression and classification) and unsupervised tasks. Even though for simpler regression tasks, such as predicting house prices simpler ML algorithms are commonly used, for more complex tasks like time series prediction and language modelling NNs are sometimes preferred. Thus, there are different structures of NNs, depending on the use case.

For regression problems, like time series prediction and language modelling recurrent neural networks (RNNs) are the most common, with some variants of them like long short-term memory networks (LSTMs) and gated recurrent units (GRUs). For classification tasks and for handling grid-like data, such as images, convolutional neural networks (CNNs) are designed. CNNs are very popular in use cases such as image recognition, object detection, and other computer vision tasks. Unsupervised neural networks, like all unsupervised ML algorithms, are trained using unlabeled data, trying to discover hidden patterns or representations. Autoencoders (AEs) are a form of unsupervised NN. Autoencoders can be leveraged for dimensionality reduction, feature and representation learning [44, 45, 46]. AEs have a varied structure with respect to the application. An autoencoder can learn a more concise depiction of the training data (i.e. incomplete autoencoder) or an overly detailed representation of the data (i.e. overcomplete autoencoder). A probabilistic extension of AEs, is the variational autoencoders (VAEs), which are typically used for data augmentation [47-50]. Other NNs that are commonly used for image synthesis are the generative adversarial networks (GANs) [51, 52].

To summarize, neural networks are cutting edge models and are highly effective in modeling complex data patterns. Their architecture can range from shallow to deep, and from fully connected to sparsely connected, tailored to specific tasks. Recurrent Neural Networks and Convolutional Neural Networks are specialized for sequential and grid-like data, respectively. Supervised neural networks are employed for tasks such as regression and classification, while unsupervised neural networks, including autoencoders, variational autoencoders, and generative adversarial networks, are utilized for uncovering hidden patterns and generating new data.

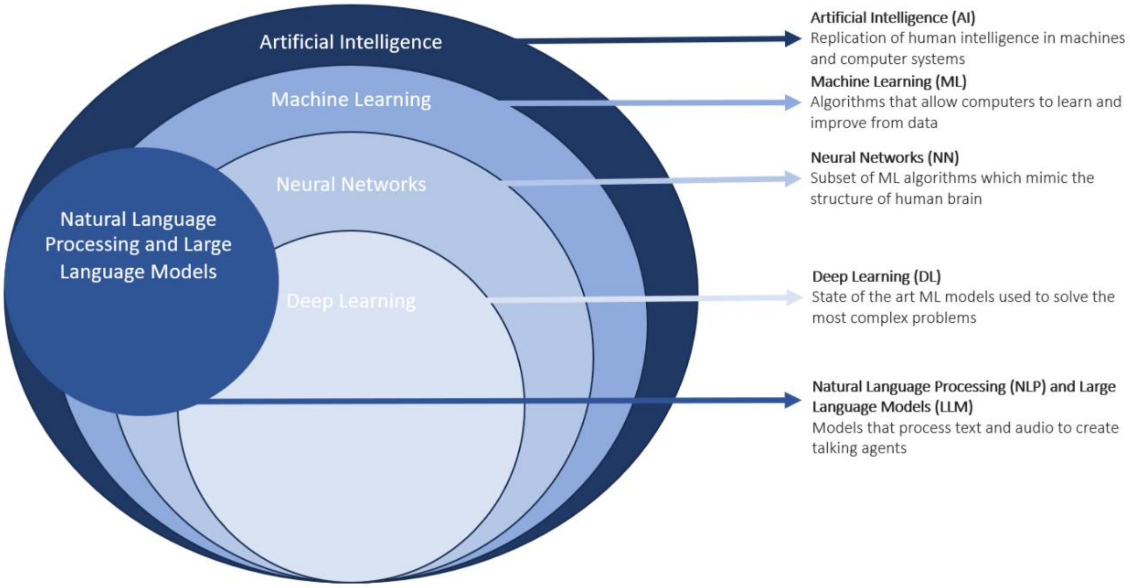

Artificial Intelligence
Machine Learning
Neural Networks
Deep Learning
Natural Language Processing and Large Language Models
Artificial Intelligence (AI)
Replication of human intelligence in machines and computer systems
Machine Learning (ML)
Algorithms that allow computers to learn and improve from data
Neural Networks (NN)
Subset of ML algorithms which mimic the structure of human brain
Deep Learning (DL)
State of the art ML models used to solve the most complex problems
Natural Language Processing (NLP) and Large Language Models (LLM)
Models that process text and audio to create talking agents

Figure 4. Comparison between Artificial intelligence, Machine learning, Neural networks, Deep learning, Natural Language Processing and Large Language Models.

A4. Sample size estimation in clinical trials and bioequivalence studies

Sample size estimation is a crucial component of clinical trials since the latter serves as the cornerstone for ensuring safety and efficacy [53]. 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 [53].

The determination of sample size can vary significantly based on the study design, outcome type, and hypothesis test specified by the investigator [54]. 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 [54]. Achieving an optimal balance between an insufficient or excessive number of participants in the sample is imperative [55]. 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 [56, 57, 58]. 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.

As with any clinical trial, similar concerns apply to bioequivalence (BE) studies, especially in cases where a generic pharmaceutical product (i.e., Test, T) is compared against the reference product (R) [57,58]. Two pharmaceutical products are deemed bioequivalent if they contain the same active substance at the same molar dose, and their equivalence is demonstrated through comparative pharmacokinetic studies, i.e., bioequivalence trials. If bioequivalence is established in the comparative pharmacokinetic trial, the two products can be considered therapeutically equivalent [57,58].

In the context of BE testing, various approaches have been proposed to address the need of recruiting a large number of volunteers, particularly for highly variable drugs [57,58]. In recent years, the emergence of in silico methods has led to the adoption of computational alternatives [56]. These computational methods are utilized for virtually increasing the sample size, a process known as data augmentation. Artificial intelligence, particularly deep learning, has proven advantageous in its several aspects of clinical studies [59]. Considering the importance of data accessibility in data-oriented and personalized healthcare, the use of AI becomes imperative. Training AI models has demonstrated considerable benefits, accelerating and simplifying every step within drug research [60,61].

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% [62,63].

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 [56]. 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 [56]. To address this issue, various methods have been proposed. Both the EMA and the U.S. FDA currently recommend using scaled BE limits [62,63]. 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 [62,63].

A5. Scope of the thesis

The aim of this thesis was to apply machine learning methods to analyze clinical data on neutralizing antibodies after COVID-19 vaccination in order to uncover any hidden relationships. Secondly, and most importantly, this thesis introduces the idea of using generative AI algorithms in clinical research. Specifically, these AI algorithms, with a particular focus on variational autoencoders, were utilized as data augmentation tools, first within clinical studies and subsequently in bioequivalence studies.

Chapter B: Materials & Methods

B1. Machine Learning

In this section, a brief introduction will be provided to cover the fundamental concepts of machine learning and deep learning algorithms. This will include an overview of key principles, methodologies, and distinctions between these two fields, highlighting how machine learning focuses on using algorithms to enable systems to learn from data, and deep learning further builds on this by using neural networks to model complex patterns and representations.

In this context, machine learning consists of a set of algorithms that analyze and find patterns within a dataset. There are two types of algorithms, supervised and unsupervised.

B1.1. Supervised

The supervised algorithms' main objective is to quantify (i.e. model) the effect of a set of independent variables $X = \{x_1, x_2, \dots, x_n\}$, on the dependent (or target) variable y and later predict the value of y, $\hat{y}$, for new unseen data. This procedure is referred as "training the model/algorithm" and it requires as input the "training data", which is a subset of the original data. Once the model is trained, its performance is evaluated on unseen data, referred as "test data". The ratio between the training and test subsets is usually 2:1. When evaluation the model, an appropriate metrics is chosen depending on the problem and the algorithm. Some algorithms require an extra step, which is called "hyper parameter tuning". At this step appropriate values for the algorithm-specific parameters must be chosen. This is achieved using the "validation data", which is also part of the original dataset. The ratio between training, validation and test data is usually 2:1:1.

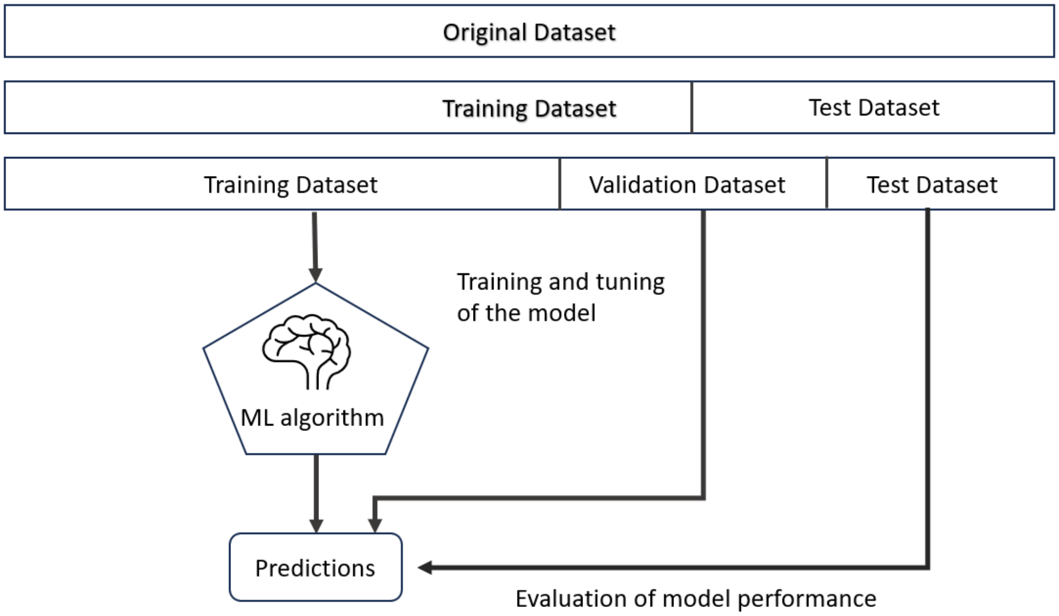


Figure 5. Process of training a supervised machine learning model. The original dataset is split into two or three parts depending on the use case. The training dataset is used to train the model, the validation dataset is used to tune the hyperparameters of the model (if any) and the test dataset is used to evaluate the model.

There are cases where the dataset does not allow for a decent validation dataset. In those cases, "k cross-validation" can be used. The training data are split into k subsets (i.e. fold). All folds have the same size. In practice, usually 5-fold cross-validation is used. Thus, the dataset L, is randomly split into 5 folds $L_k = \{L_1, L_2, \dots, L_5\}$ where each $L_k, k = 1, 2, \dots, 5$ contains 20% of the training data. Later, the model is trained five times, each time using four out of five subsets for training and the remaining one of validation. That is model f_1, will be trained using $\{L_2, \dots, L_5\}$ and L_1, for validation, model f_2, will be trained using $\{L_1, L_3, \dots, L_5\}$ and L_2, for validation and so on. In the end the models' performance are averaged to get the final model performance [64].

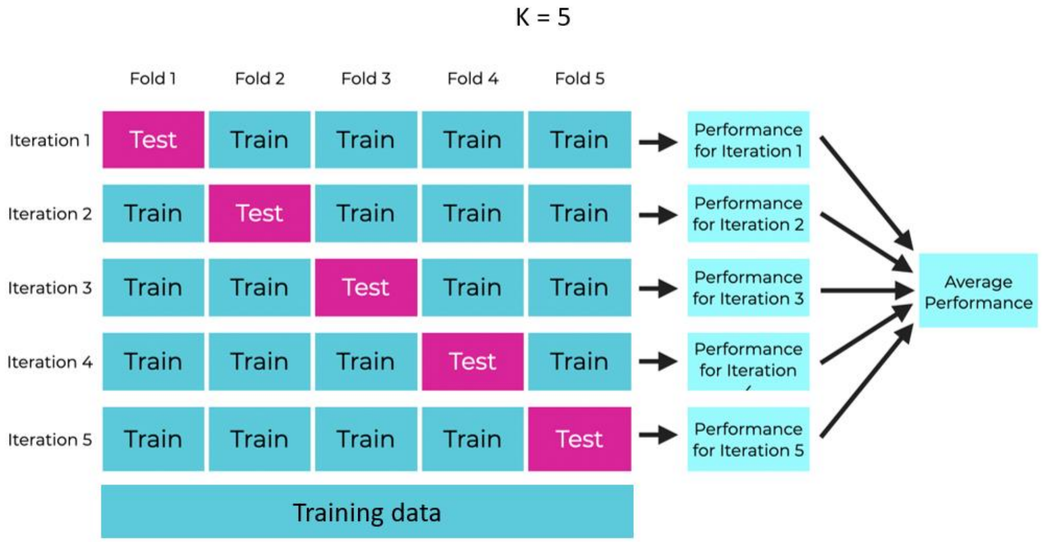


Figure 6. Cross validation for k=5 folds. The dataset is divided into five equal parts. The model is then trained and validated five times. Each time, one of the five parts is used as the validation set, while the remaining four parts are used for training. This process is repeated five times, ensuring that each part of the dataset is used exactly once as the validation set. After all iterations, the performance metrics from each of the five runs are averaged to give a final performance estimate. Adapted from [65].

The goal of all supervised machine learning algorithms is to be as accurate as possible when predicting. We can evaluate the predictions by assessing the magnitude of the error term, $error = y - \hat{y}$. An issue that often arises is “overfitting”. When a model is too complex it captures the noise and details when been trained, which affects negatively its performance on new unseen data. Thus, the model predicts well (i.e. low training error) in the training data, but poorly (i.e. high-test error) in the test data.

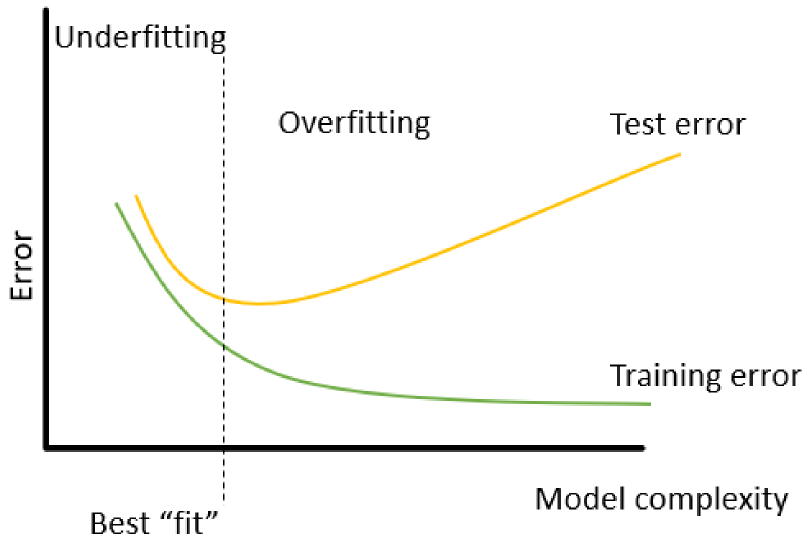


Figure 7. Overfitting model showcase. The error in the training dataset will always decrease whereas the error in the test data will start to increase while the model become more and more complex. The optimal model complexity is when the error in the test dataset just begins to increase.

B1.1.1. Regression

Regression in supervised machine learning refers to use cases where the target variable y is continuous. Predicting the value of bitcoin, the energy consumption of a house, the sales of a brand or the election results are regression problems. In all regression machine learning problems, we have a collection of labelled examples $\{(x_i, y_i)\}_{i=1}^{N}$, where N is the number of observations (i.e. examples) in the dataset, x_i, is a vector of features of dimension D, of example $i = 1, \dots, N, y_i \in \mathbb{R}$ is the target variable and $x_i^j \in \mathbb{R}, \mathrm{j} = 1, \dots, \mathrm{D}$.

Mean squared error and mean absolute error (MAE) are the most common metrics used to evaluate the result of a supervised regression ML algorithm. Consider $\widehat{y_i}$ *being the prediction of* y_i, for the i-th example MSE and MAE are defined as $\frac{1}{N}\sum_{i=1,\dots,N}(y_i - \widehat{y_i})^2$ and $\frac{1}{N}\sum_{i=1,\dots,N} |y_i - \widehat{y_i}|$ respectively.

B1.1.1.1. Linear regression (and Polynomial and Ridge Regression)

An algorithm that learns a model by combining the input features in a linear manner is linear regression. Thus, linear regression can be expressed as the model f:

$$f_{w,b}(x) = wx + b \tag{1}$$

where w is a D-dimensional vector usually referred to as "weights" and $b \in \mathbb{R}$, usually referred to as "bias" term. The model f, is parametrized by the (w, b), and is used to make predictions of the y (i.e. $\hat{y}$). Different parametrized values (w, b), will produce different predictions. Finding the optimal (w, b) (i.e. (w^*, b^*)), which will make the predictions most accurate, is the goal.

Let the $\widehat{y_i}$, be the prediction for Y on the i-th example. Then the $e_i = y_i - \widehat{y_i} = y_i - f(x_i)$, represents the i-th residual. We define the residual sum of squares (RSS) as:

$$RSS = e_1^2 + \cdots + e_N^2 \tag{2}$$

and the loss function as:

$$(y_i - f(x_i))^2 \tag{3}$$

The goal is to find the optimal (w^*, b^*), that minimize the average loss function. Thus,

$$\min_{w,b} \frac{1}{N} \sum_{i=1,\dots,N} (y_i - f(x_i))^2 \tag{3}$$

The optimal (w^*, b^*), can be analytically computed and it is shown to be [66]

$$w^* = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{4}$$

$$b^* = \bar{y} - w\bar{x} \tag{5}$$

where $\bar{y}, \bar{x}$ are the means of the target variable and the input features respectively.

Linear regression can be extended in multiple ways. One can extend the relationship between x and y to be a polynomial of n degree.

The polynomial regression equation is:

$$y = w_0 + w_1 x + w_2 x^2 + w_n x^n \tag{6}$$

where $w_0, w_1, \dots, w_n$ are the coefficients of the polynomial. The polynomial regression allows for more complex relationships, whereas linear regression is a simple case of polynomial regression where $n = 1$ [66].

Another extension of linear regression is "ridge regression", which addresses multicollinearity (when independent variables are highly correlated) and overfitting. It adds a regularization term to the cost function to penalize large coefficients.

The ridge regression equation is:

$$J(w,b) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (wx_i + b))^2 + \lambda\sum_{j=1}^{D} w_j^2 \tag{7}$$

where λ is called "regularization" parameter, that controls the amount of shrinkage applied to the coefficients [66]. Ridge regression, also known as Tikhonov regularization.

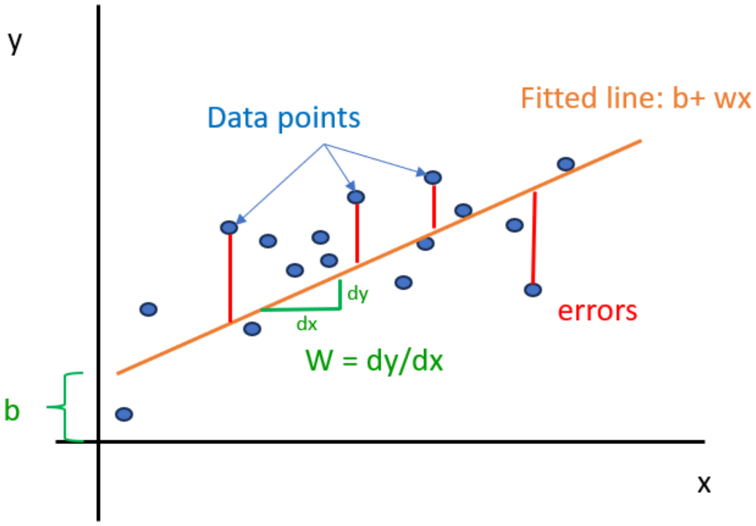


Figure 8. Line fitted to data using linear regression. The "b" is the intercept of the line with the y-axis and "w" is the slope of the line. The distance of each datapoint to the line, is the "error" for this datapoint.

B1.1.1.2. Decision Trees and Random Forest

Decision tree is a directed graph which includes nodes that signify decisions, branches that depict decisions' outcomes, and leaf nodes that represent the final outcomes or predictions. The flow has a top-down approach, starting from the root node which represents the full dataset, passing through the internal nodes that correspond to the decisions and ending in the leaf nodes that represent the final outcomes.

Thus, given training feature $x_i \in R^n, i = 1, \dots, l$ and a target vector $y \in R^l$, a decision tree divides the feature space recursively, grouping samples with similar target values together.

Let Q_m be the data at node m with n_m samples. For each candidate split $\mathrm{s} = (i, p_m)$ consisting of a feature i and threshold p_m, partition the data into $Q_m^{\text{left}}(s)\ and\ Q_m^{\text{right}}(s)$ subsets:

$$Q_m^{\text{left}}(\mathrm{s}) = (x, y)/x_i \leq p_m \tag{8}$$

$$Q_m^{\text{right}}(\mathrm{s}) = Q_m\ Q_m^{\text{left}}(\mathrm{s}) \tag{9}$$

The “goodness of split” of node m is then computed:

$$G(Q_m, \mathrm{s}) = \frac{n_m^{left}}{n_m} H\left(Q_m^{left}(\mathrm{s})\right) + \frac{n_m^{right}}{n_m} H\left(Q_m^{right}(\mathrm{s})\right) \tag{10}$$

H is a loss function which depends on the task being solved.

The goal is to minimize the impurity:

$$\mathrm{s}^* = argmin_{\theta} G(Q_m, \mathrm{s}) \tag{11}$$

$Q_m^{\text{left}}(s^*)$ and $Q_m^{\text{right}}(s^*)$ are recursed, until $n_m < min_{samples}$ or $n_m = 1$ (i.e. maximum depth allowed is reached).

For regression problems it is common to minimize MSE:

$$\overline{y_m} = \frac{1}{n_m} \sum_{y \in Q_m} y \tag{12}$$

$$H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} (y - \overline{y_m})^2 \tag{13}$$

as shown in [68].

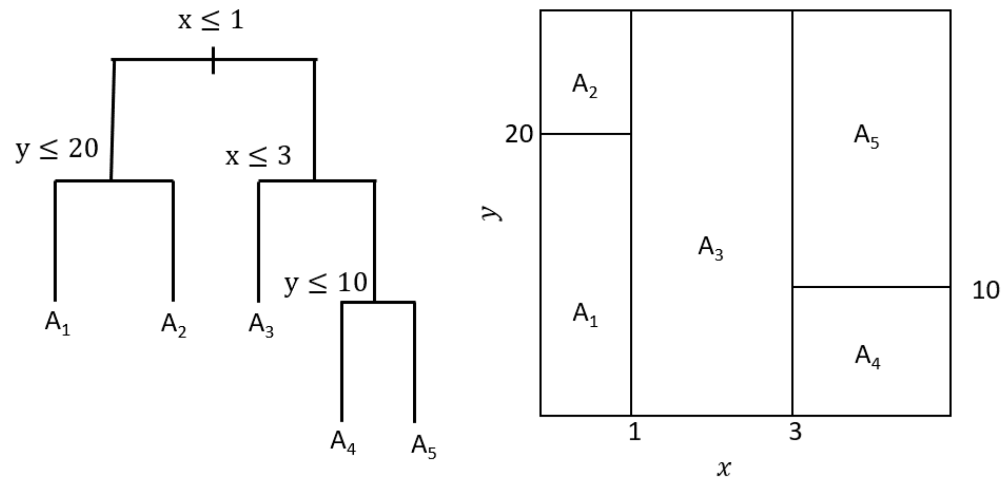


Figure 9. A decision tree (left) corresponding to the partition of the 2D space (right).

Random forest combines multiple decision trees, thus is an ensemble method. This is done in a “bagging” way. In bagging, the machine learning algorithm, combines the predictions of multiple weak learners (in this case, decision trees) trained on different subsets of training data.

This achieves reducing the correlation between trees [64].

Random forest pseudo-algorithm as presented in [69]

```
1. For k = 1 to K:
    a. Generate a random bootstrapped sample of size $N$ from the training dataset.
    b. Create a tree $T_k$, using the data from step (a) by repeating:
        i. Randomly choose $m$ out of the $d$ variables.
        ii. Choose the optimal split from the $m$.
        iii. Split the parent node into 2 nodes.
        iv. Repeat steps i, ii, iii until the minimum node size $min_{samples}$ is reached.
2. Output the ensemble of trees $T_k{}_{k=1}^{K}$.
```

To make a prediction at a new point x for regression:

$$\widehat{f_{x_\iota}^{K}}(x) = \frac{1}{K}\sum_{k=1}^{K} T_K(x) \tag{14}$$

Where K is the total number of trees.

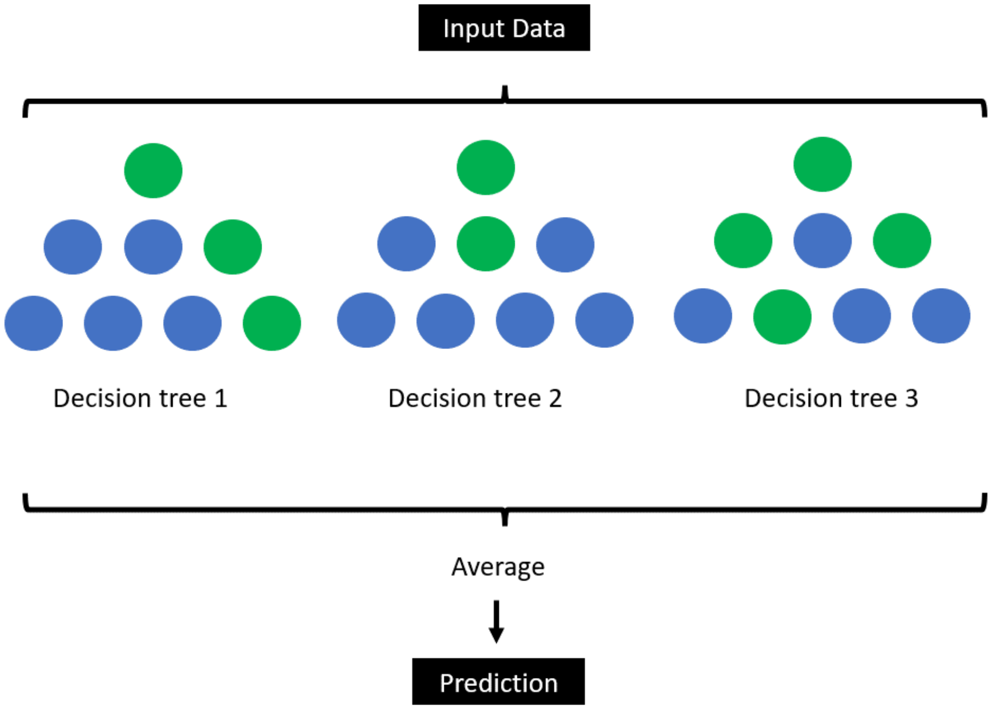


Figure 10. Random forest averaging the result of multiple decision trees for prediction in regression use cases.

B1.1.1.3. Gradient Boosting Machines

Boosting involves repeatedly creating multiple models from the original training data using a weak learner. Each successive model is designed to correct the errors of its predecessor. The ultimate ensemble model is formed by iteratively combining several weak models in a specific manner [64]. The most common boosting algorithm is gradient boosting machine (GBM).

Mathematically, the model at iteration t, can be described as follows [69]:

$$F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x) \tag{15}$$

where $F_t(x)$ is the model at iteration t , η is the learning rate, and $h_t(x)$ is the new tree fitted to the pseudo-residuals.

For regression tasks, the following loss function is being minimized:

$$L\big(y, F(x)\big) = \frac{1}{2}\big(y - F(x)\big)^2 \tag{16}$$

Which is known as MSE.

The sum of the predictions from all the trees is the final prediction:

$$\hat{y} = \sum_{m=1}^{m} f_m(x) \tag{17}$$

There are various alternatives of GBM, such as extreme gradient boosting (XGBoost) and light gradient boosting (LightGBM). The first is an optimized implementation of GBM for speed and performance, which includes also regularization to prevent overfitting [70], whereas the latter is optimized for handling large datasets with low memory usage [71].

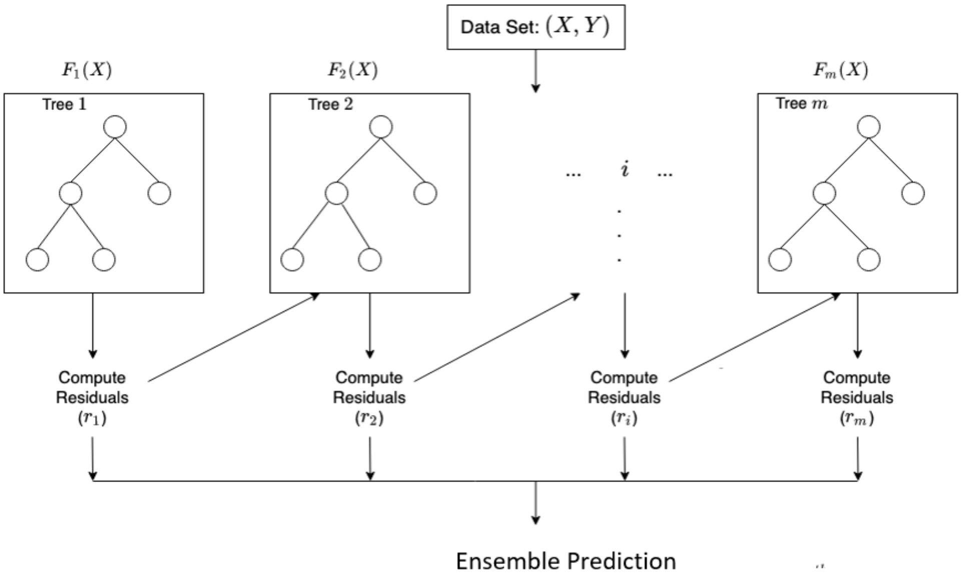

Data Set: (X, Y)
$F_1(X)$
Tree 1
$F_2(X)$
Tree 2
... i ...
$F_m(X)$
Tree m
Compute Residuals (r_1)
Compute Residuals (r_2)
Compute Residuals (r_i)
Compute Residuals (r_m)
Ensemble Prediction

Figure 11. A gradient boosting model combining multiple weaker learners (decision trees) to make as accurate prediction as possible. Adapted from [72].

B1.1.1.4. Support Vector Regression

The core concept of Support Vector Regression (SVR) is to identify a function that closely estimates the relationship between the input variables and the target variable. This function aims to predict values within a specified margin of tolerance (ϵ), ensuring that the predictions are as accurate as possible while allowing for some flexibility [73].

In other words, SVR is finding a function $f(x)$ that is flat as possible and is at maximum ϵ-far from the actual target values y_i for all training data. The function $f(x)$ is typically a linear function within the feature space, represented as:

$$f(x) = \langle w, x \rangle + b \tag{18}$$

where $\langle w, x \rangle$ is the dot product of the weight vector w and the input vector x, and b is the bias term.

As explained in [73], SVR is minimizing the following cost function:

$$\frac{1}{2}|w|^2 + C\sum_{i=1}^{n} L_\epsilon\big(y_i, f(x_i)\big) \tag{19}$$

where $|w|^2$ ensures that the function is flat, C is a regularization parameter that regulates the balance between the smoothness of the function and the extent to which deviations larger than ϵ are tolerated, and $L_\epsilon\big(y_i, f(x_i)\big)$ is the ϵ-insensitive loss function defined as:

$$L_\epsilon\big(y_i, f(x_i)\big) = \max(0, |y_i - f(x_i)| - \epsilon) \tag{20}$$

This loss function means that errors less than ϵ are ignored, and only errors greater than ϵ contribute to the cost.

SVM regression conducts linear regression in the high-dimensional feature space using the ϵ-insensitive loss function and while also simplifies the model by minimizing $|w|^2$. This is equivalent by minimizing the following function:

$$\frac{1}{2}|w|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{21}$$

subject to:

$$\begin{cases} y_i - f(x_i) \le \epsilon + \xi_i^* \\ f(x_i) - y_i \le \epsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0, \mathrm{i} = 1, \ldots, \mathrm{n} \end{cases} \tag{22}$$

where ξ_i, ξ_i^* are slack variables that allow some errors in predictions.

As described in [73], to solve the optimization problem, SVR uses Lagrange multipliers and the dual formulation, similar to SVM. The dual problem is given by:

$$\max_{\alpha,\alpha^*}\left\{-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)\big(\alpha_j - \alpha_j^*\big)K\big(x_i, x_j\big) - \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*)\right\} \tag{23}$$

subject to the constraints:

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \tag{24}$$

$$0 \leq \alpha_i, \alpha_i^* \leq C \tag{25}$$

where α_i and $\alpha_i^*$are the Lagrange multipliers, and $K(x_i, x_j)$ is the kernel function. In the simpler case of the linear kernel $K(x_i, x_j) = \langle x_i, x \rangle$.

In this case the final regression formula is:

$$\sum_{i=1}^{n}(\alpha_i \ - \ \alpha_i^*)\,\langle x_i, x \rangle + b \tag{26}$$

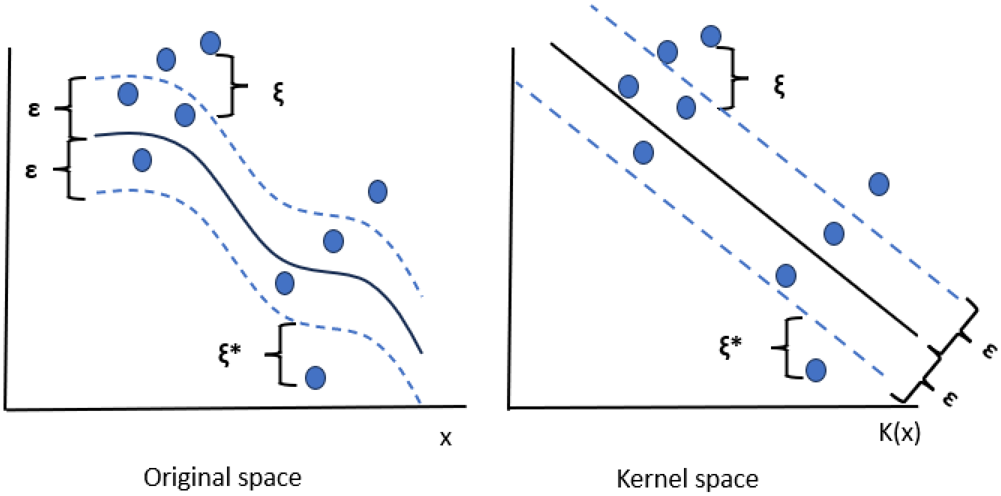


Figure 12. Support vector regression with kernel. The original space (left) is mapped to the Kernel space (right) using a kernel function "K". The fitted line (black) is then transformed from a curve to linear.

B1.1.2. Classification

The difference between regression and classification is the nature of the target vector Y. In classification y_i is categorical, meaning that it has a fixed amount of possible values. If the amount of possible values equals to 2, then the classification is called "binary" classification, otherwise it is called "multi-class". Examples of binary classification are predicting if an image is a cat or a dog, or if a person will develop cancer or not, whereas an example of multi-class classification is predicting which country will win the most medals in Olympic games.

The most common method for evaluating the performance of a supervised classification machine learning algorithm is the confusion matrix. For classes $C = \{C_k\}_{k=1}^{K}$, the confusion matrix is a KxK table, where each row represents the instances in actual class and each column the instances in the predicted class, thus the element $C_{i,j}$ represents the instances that belong to class i but were predicted to be in class j.

The diagonal components of the confusion matrix, $C_{i,i}$ are called true positives (TP) and represent the number of instances that were predicted correctly as class i. The off diagonal elements in the column j, without the diagonal, represent the instances that were incorrectly predicted as class j and are called false positives (FP). The off diagonal elements in the row i, without the diagonal, represent the instances that were incorrectly predicted as not being in the class i and are called false positives (FN). Lastly the true negatives (TN) are the instances correctly predicted as not being in particular class.

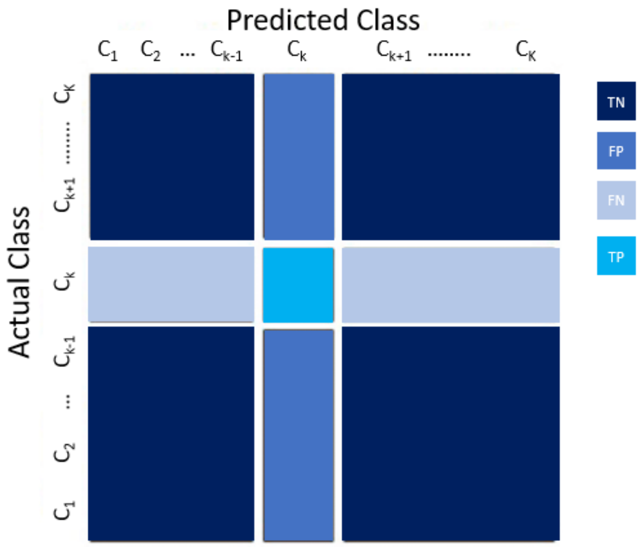


Figure 13. Confusion matrix for K classes.

From the confusion matrix there are a lot metrics that can be calculated to evaluate the performance of the model for each class i. The most common are:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{27}$$

$$Recall_1 = \frac{TP_i}{TP_i + FN_i} \tag{28}$$

$$Accuracy_i = \frac{TP_i + TN_i}{TP_i + TN_i + FN_i + FP_i} \tag{29}$$

$$F1\ Score_i = 2\frac{Precision_i * \ Recall_i}{Precision_i \ + \ Recall_i} \tag{30}$$

Each of the above metrics is used depending on the use case. Accuracy is used when the classes are balanced, that is the number of examples at each class is similar, precision is used when the costs of false positives is high whereas recall is used when the cost of false negatives is high.

B1.1.2.1. Logistic regression

Logistic regression is a classification algorithm. It aims to model the posterior probabilities of the K classes as a linear function of x_i, while preserving the probabilities properties that they sum to 1 and their range is in $[0,1]$.

The model for the K classes is:

$$\mathrm{Pr(G = K|X = x)} = \frac{1}{1 \ + \ \sum_{l=1}^{K-1} e^{wx+b}} \tag{31}$$

Notice that $\sum_{l=1}^{K} Pr(G = K|X = x) = 1$ and with the simple binary classification case, where $K = 2$, (26) equals the sigmoid function.

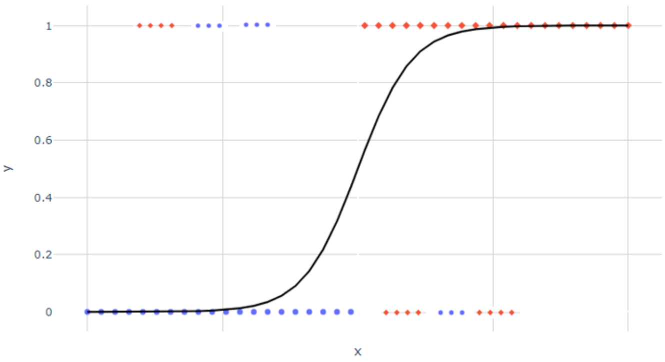


Figure 14. The sigmoid function fitted to binary data.

Similarly to logistic regression we are looking for the optimal parameters (w^*, b^*). In the simply binary case, the optimal parameters are obtained by maximizing the likelihood function:

$$max_{w,b}L_{w,b}, \tag{32}$$

where

$$L_{w,b} \overset{\text{def}}{=} \prod_{i=1,\dots,N} f_{w,b}(x)^{y_i}(1 - f_{w,b}(x))^{(1-y_{i)}} \tag{33}$$

Unlike to linear regression, there is no analytical solution for (29), and the most common optimization procedure to solve (29) is gradient decent [64].

B1.1.2.2. Decision Trees and Random Forest

Decision trees is a non-parametric algorithm that can be used for classification, meaning it does not make any assumption regarding the underlying distribution of the data. The trees are built recursively, splitting the dataset based on the feature that provides the "best" split. The best split is defined by an impurity measure.

One common impurity measure is the Gini impurity index:

$$Gini = 1 - \sum_{j} (p(j|t))^2 \tag{34}$$

where $p(j|t)$ is the relative frequency of class j at node t. Gini ranges from 0, when all records belong to one class, implying most interesting information, to $1 - 1/K$, when records are equally distributed among all classes (K number of classes) implying least interesting information [73].

Another common impurity measure is entropy:

$$Entropy = -\sum_{i=1}^{n} p(j|t) \log_2\big(p(j|t)\big) \tag{35}$$

where $p(j|t)$ is the relative frequency of class j at node t. Entropy measures the homogeneity of a node. Its values range from 0, when all records belong to one class, implying most information, to $\log_2 K$, when records are equally distributed among all classes implying least information [73].

Another common measure is Information Gain:

$$GAIN = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right) \tag{36}$$

where parent node p, is split into k partitions and n_i is the number of records in partition i.

Information Gain measures the reduction in entropy achieved because of the split. Chooses the split that achieves most reduction [73].

A common loss function is the cross-entropy, which measures the difference between the predicted probabilities and the actual class labels:

$$Crossentropy = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} log\big(p_{i,k}\big) \tag{37}$$

where N is the number of examples, K is the number of classes, $y_{i,k}$ is binary, indicating if k is the correct class label for the example I, and $p_{i,k}$ is the predicted probability of example i belonging in class k.

In some cases, where overfitting is observed, “pruning” is used. Pruning is a technique that reduces the size of a tree by removing branches that have little importance and do not provide a lot of improvement in terms of accuracy.

Same as linear regression, a lot of decision trees combined, they form a random forest.

Like in regression, the pseudo-algorithm as presented in [69]:

```
1. For k = 1 to K:
    a. Generate a random bootstrapped sample of size $N$ from the training dataset.
    b. Create a tree $T_k$, using the data from step (a) by repeating:
        i. Randomly choose $m$ out of the $d$ variables.
        ii. Choose the optimal split from the $m$.
        iii. Split the parent node into 2 nodes.
        iv. Repeat steps i, ii, iii until the minimum node size $min_{samples}$ is reached.
2. Output the ensemble of trees $T_{k}{}_{k=1}^{K}$.
```

To make a prediction at a new point x for regression: Let $\widehat{C_k}(x)$ be the class prediction of the k-th random-forest tree. Then $\widehat{C_{rf}^{K}}(x) \; = \; majority\ vote\ \{\,\hat{C}_k(x)\}_1^K$.

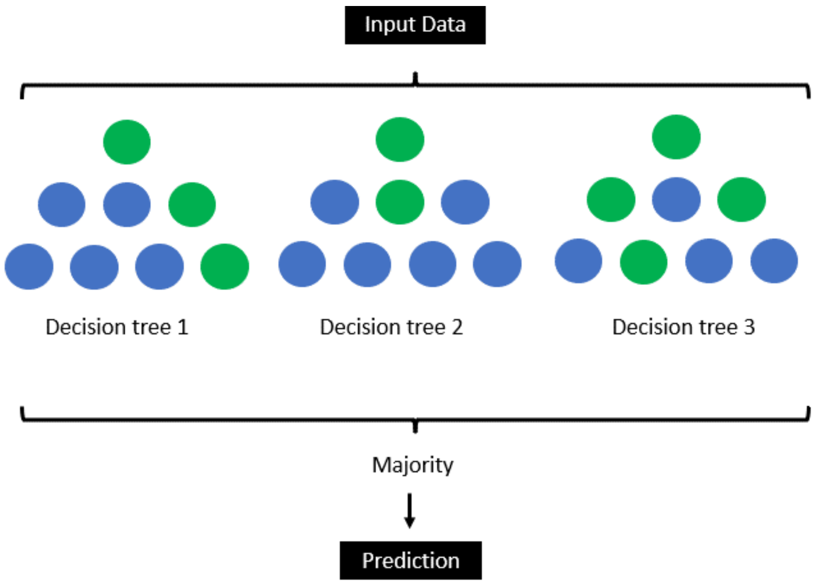


Figure 15. Random forest combining multiple decision trees for classification using the majority rule.

B1.1.2.3. Gradient Boosting Machines

GBM in classification problems is very similar with GBM in the regression. The difference lays in the loss function used. The logistic loss is used for binary classification:

$$L(y,F(x)) = -\frac{1}{n}\sum_{i=1}^{N}[y_i \log(F(x)) \; + \; (1-y_i)log(1-F(x))] \tag{38}$$

Whereas for multiclass classification the softmax is often used:

$$L\big(y,F(x)\big) == -\frac{1}{n}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\, log\big(F(x)\big) \tag{39}$$

In binary classification the prediction is done by using a threshold:

$$\hat{y} = \begin{cases} 1 if \sigma\left(\sum_{m=1}^{M} f_m(x)\right) > 0.5 \\ 0 \quad otherwise \end{cases} \tag{40}$$

where σ is the sigmoid function, whereas in the multiclass by taking the argmax of the summed probabilities:

$$\hat{y} = \underset{k}{\operatorname{argmax}}(\sum_{m=1}^{M} f_{mk}(x)) \tag{41}$$

where $f_{mk}(x)$ is the prediction of class k from the m tree.

B1.1.2.4. Support Vector Machines

Support vector machines (SVM) aims to find a hyperplane that separates the data into different classes as good as possible, meaning that the distance of the boarders of the classes is as big as possible. There are cases where the data are not linearly separable. Figure 16, shows non-linearly separable data points of 2 classes in 2 dimensions (left), but if we transform these data into 3 dimensions then they become linearly separable (right).

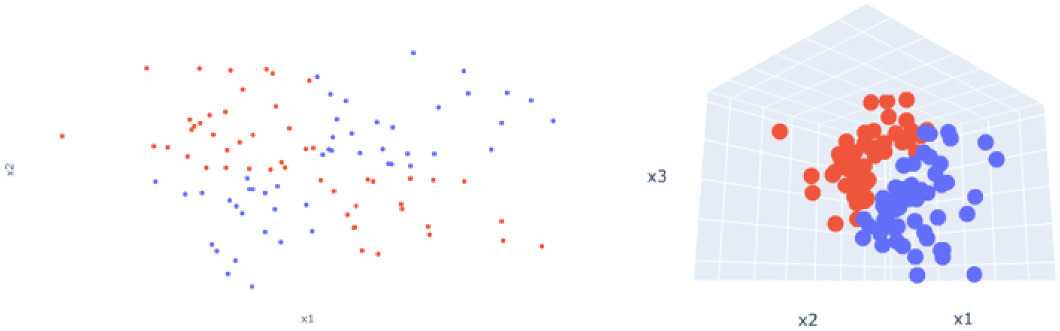


Figure 16. Non-linearly separable data of 2 classes in 2D (left) which become linearly separable when they are transformed into 3D (right).

The above technique is commonly referred as the “kernel trick”. As shown in Figure 17, there can be multiple hyperplanes that separate the data.

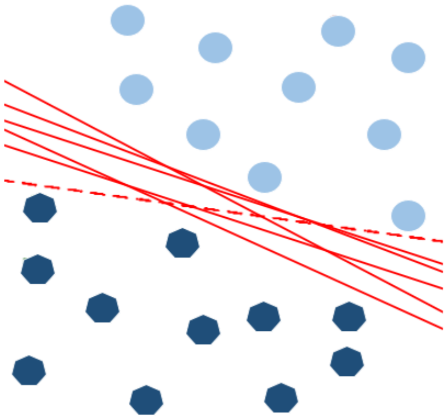


Figure 17. Possible ways to linearly separate the data of the 2 classes in 2D.

The optimal hyperplane that SVM is searching for, is the one that maximizes the margin, as shown in Figure 18.

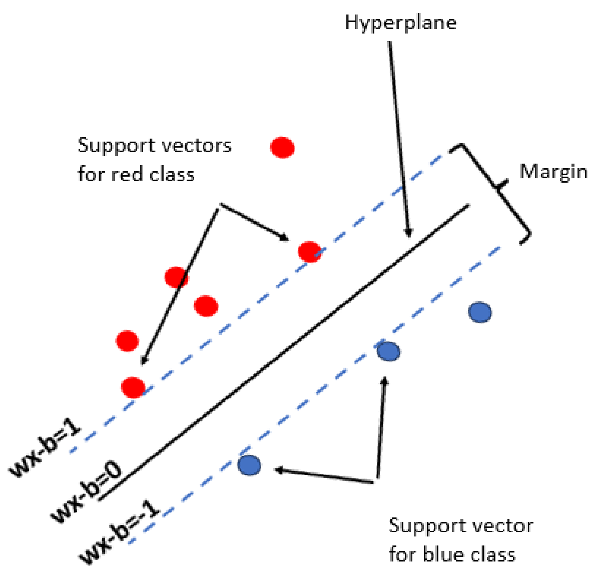


Figure 18. Optimal hyperplane that maximizes the margin between the 2 classes.

As shown in [73] the objective function for SVM is:

$$\begin{cases} \min_{w,b,\xi} \frac{1}{2} |w|^2 + C \sum_{i=1}^{n} \xi_i \\ subject\ to \\ y_i(wx_i + b) \geq 1 - \xi_i \\ and \\ \xi_i \geq 0 \end{cases} \tag{42}$$

Where w is the weight vector, b is the bias term, x_i are the feature variables, C is the regularization parameter, y_i are the class labels and ξ_i are "slack variables" for the i-th observation which allows some flexibility, especially when the data are not linearly separable [73].

B1.1.2.5. K-nearest neighbors

Another supervised machine learning algorithm is k-Nearest Neighbors (kNN). One main difference with the rest ML algorithms, is that kNN does not disregard the training data once trained. The training data are saved in memory, and are utilized by kNN to find the k closest examples of a new, previously unseen example, using a D-dimensional distance measure, and for the classification it returns the majority class of the k examples.

The most common distance measures used in kNN are the Euclidean distance:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_D - q_D)^2} \tag{43}$$

where p, q are two points in D-dimensional space.

Another common distance metric is the cosine similarity:

$$\cos(p,q) = \frac{p \cdot q}{|p||q|} = \frac{\sum_{i=1}^{D} p_i q_i}{\sqrt{\sum_{i=1}^{D} p_i^2}\sqrt{\sum_{i=1}^{D} q_i^2}} \tag{44}$$

where p, q are two points in D-dimensional space, $p = (p_1, p_2, \dots, p_D)$ and $q = (q_1, q_2, \dots, q_D)$.

The pseudo algorithm of kNN when the Euclidean distance is used as distance metric, is as follows:

```
1. Input the dataset (split by training and test).
2. Use the optimal $k^*$ (this needs to be predefined by the researcher).
3. For each point in test data:
      i. Find the Euclidean distance with respect to all training data points.
     ii. Choose the first $k^*$ points.
    iii. Assign to the test data point, the majority class.
```

In the aforementioned pseudo algorithm, it is evident that the researcher needs to find the optimal k (k^*). This is usually achieved with a brute force approach. Assuming that the set of candidate k is $k_{candidate} = \{1,2,3,\dots k_{max}\}$, then for each $k \in k_{candidate}$, the aforementioned algorithm is ran, and the accuracy is calculated.

Following the Occam's razor also knows as principle of parsimony, which suggests that the simplest explanation is usually the correct one, k^* is the minimum k for which a desirable performance is achieved.

The choice of k^*, has a significant role in the performance of the algorithm. Figure 19 shows that the new data point (green), will be assigned to the red class if $k = 3$ whereas if $k = 5$ it will be assigned to the blue class.

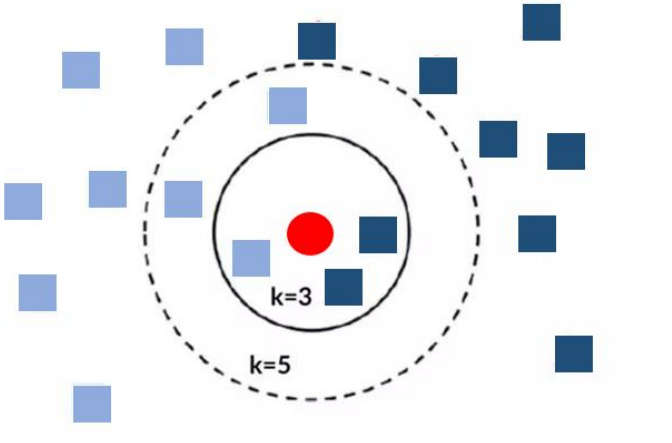


Figure 19. K-nearest-neighbors example. The new data point (circle) will be assigned to the dark blue class if k=3 whereas if k=5 it will be assigned to the light blue class.

B1.1.3. Feature importance

Feature importance in classification refers to techniques used to determine which feature has the most influence on the outcome. All the classification algorithms except KNN, have their own feature importance technique.

In linear models, like logistic regression and SVM, the feature importance is defined as:

$$\text{Importance}(x_i) = \mid \beta_i \mid \tag{45}$$

where β_i is the coefficient of variable x_i.

In models that are tree-based, like decision trees and random forests the importance is measured using GINI importance, which measures the total decrease in node impurity brought by a feature across all trees in the forest (i.e. mean decrease in impurity):

$$Gini\ Importance = \sum_{t=1}^{T} \sum_{n \in N_t} \Delta I(n, f) \tag{46}$$

where T is the total number of trees, N_t is the set of nodes in tree t, and $\Delta I(n, f)$ is the decrease in impurity at node n due to feature f [69].

Lastly for boosting methods, like GBM, the importance feature is similar to (46):

$$Importance = \sum_{t=1}^{T} \sum_{n \in N_t} \Delta L(n, f) \tag{47}$$

where $\Delta L(n, f)$ is the decrease in loss at node n due to feature f [69].

B1.2. Unsupervised

Unsupervised machine learning is a type of ML where the model is trained using data that don't have labels. The goal of this type of ML algorithms is to find hidden patterns or structures in the input data. Unsupervised machine learning encompasses dimensionality reduction and clustering techniques.

During clustering data points are grouped into segments based on a similarity metric, whereas in dimensionality reduction, the original number of features is reduced, by creating new features from the existing ones while at the same time preserving as much information as possible.

The advantage of this type of algorithms is that the unlabeled data are more accessible, though the absence of labels makes it hard to evaluate the outcome of the model thus making the results less interpretable.

B1.2.1. K-means

A common method for clustering involves defining a cost function over a set of possible clusterings with parameters. The optimal partitioning is the one that minimizes the cost [76]. K-means is one of the most common machine learning algorithms used for clustering that follows this principle to partition the dataset into k distinct, non-overlapping clusters.

K-means partitions the dataset into disjoin sets $C_1, C_2, \dots, C_k$, where to each C_i corresponds a μ_i, which is called centroid. Each datapoint in C_i belongs to X_i, which is a metric space and μ_i lives in that space. The k-means objective function measures the squared distance between each point q_i that belongs in C_i, from its corresponding centroid μ_i. The centroid μ_i is defined as:

$$\mu_i(C_i) = argmin \sum_{x \in C_i} d(x, \mu)^2 \tag{48}$$

thus k-means objective is:

$$\min_{\mu_1, \mu_2, \dots, \mu_k \in X} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2 \tag{49}$$

The k-means algorithm as described in [76] is as follows:

1. Having as input the space $X \subset \mathbb{R}^n$ and the number of the desired clusters k.
2. Initialize random k centroids$\mu_1, \mu_2, \ldots, \mu_k$.
3. Repeat until convergence:
    a. $\forall i \in |k|\ set\ C_i = x \in X : i = argmin_j \|x - \mu_j\|$.
    b. $\forall i \in |k|\ update\ \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$.

The afore mentioned procedure is shown in Figure 20:

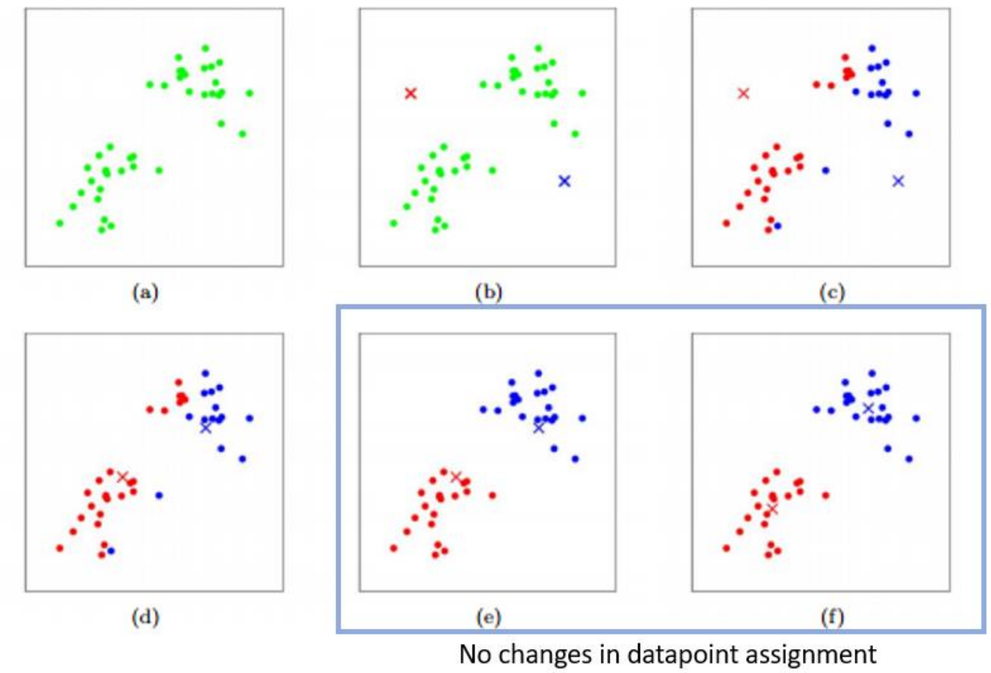


Figure 20. K-means algorithm example for k=2. The dots are the data points, and the x are the centroids. (a) shows the original dataset, (b) the random initialization of the centroids, (c-f) the k-means iterations until convergence is achieved.

Determining the ideal number of clusters, k is usually achieved via the elbow method. The elbow method performs k-means clustering across a spectrum of values of k, usually 1 to 10. Then for value of k, the sum of squared distances between data points and their corresponding cluster centroids is calculated. This is also known as the within-cluster sum of squares (WCSS). Finally, a plot is created where the x-axis is the value of k and the y-axis is the corresponding WCSS, and the elbow point is chosen, as shown in Figure 21.

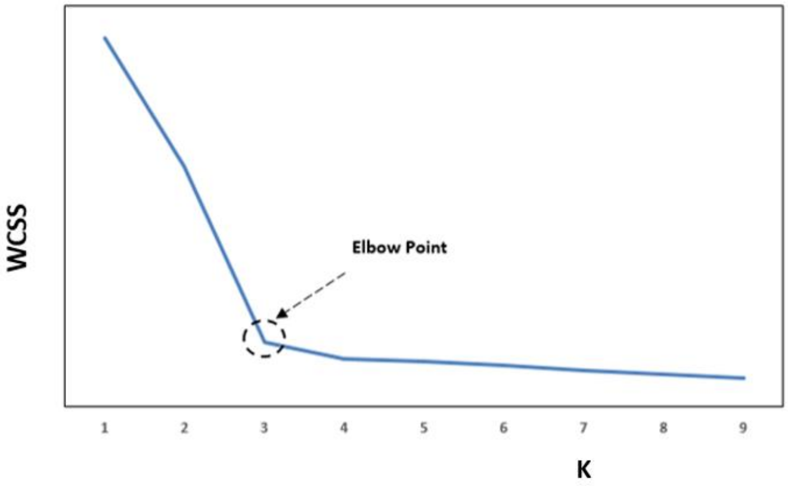


Figure 21. The elbow plot and elbow point for k-means iterations with k ranging from 1 to 9. On the y-axis is the within clusters sum of squares (WCSS) and on the x-axis is the possible k. The larger the k the smallest the WCSS. The elbow point, is where the rate of decrease of the WCSS flattens, resembling an elbow shape.

B1.2.2. Hierarchical Clustering

The outcome of the K-means algorithm is significantly influenced by the predefined number of clusters as well as the initial positioning of data points. These factors play a crucial role in determining the convergence and final configuration of clusters, as variations in either can lead to differing partitioning results. Consequently, it is necessary to predefine the desired number of clusters and establish the initial positions from which the clustering process will commence.

Hierarchical clustering operates in a distinct manner, as it does not require pre-specification of the number of clusters. Instead, it necessitates the definition of a dissimilarity metric between groups of observations. This metric is derived from the pairwise dissimilarities among observations within the groups under comparison, allowing for the formation of a nested clustering structure without a predefined cluster count. Hierarchical clustering builds a hierarchy of clusters. At each level of this hierarchy, clusters are formed by merging smaller clusters from the level below. Initially, at the lowest level, each cluster consists of just one observation. Up the hierarchy, clusters are progressively merged until, at the highest level, all observations are clustered into one group.

Hierarchical clustering is split into two categories. Agglomerative, where at the beginning each cluster contains a single data point, and later the closest pairs are merged together until only one cluster is left, commonly referred as “bottom-up” approach, and divisive, usually referred as “top-down”, begins with a single cluster which contains all data, and is later split into smaller groups until all groups contain a single datapoint [69]. The result of hierarchical cluster is a dendrogram as shown in Figure 22.

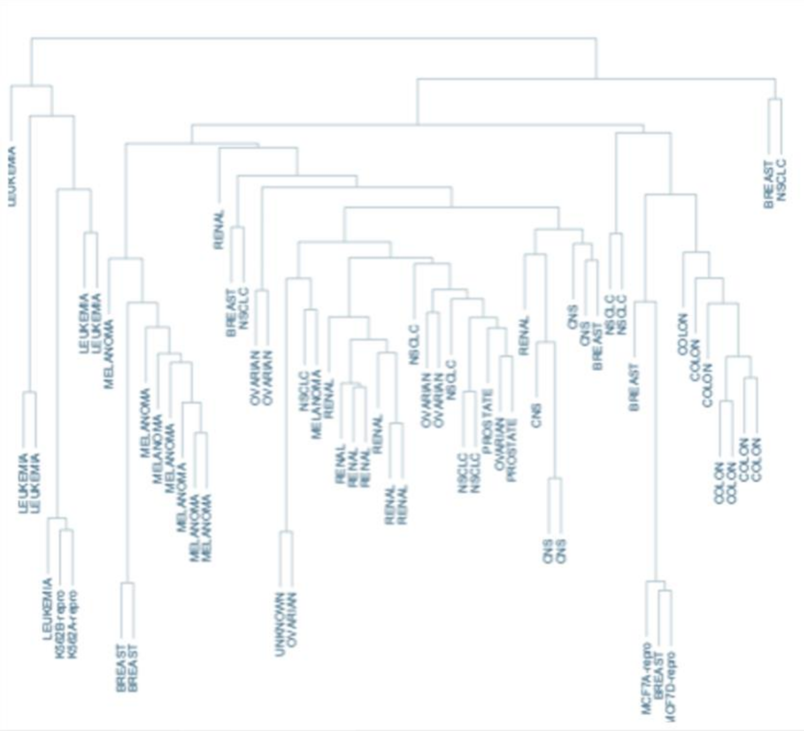


Figure 22. Agglomerative hierarchical clustering on human tumor microarray data. At the beginning each datapoint belongs to a single cluster and at each step, the closest datapoints are merged together.

During agglomerative clustering of N data points, at each $N-1$ steps a metric of dissimilarity between two groups (clusters) must be defined. Assuming two of these groups, A and B, the dissimilarity $d(A,B)$ is calculated from the set of pairwise dissimilarities of the observations d_{ij}, where $i \in A$ and $j \in B$, where d_{ij} can be defined as (43) or (44). There are different ways defining $d(A,B)$. Single linkage takes the smallest distance between a pair:

$$d_{SL}(A,B) = \min_{i \in A \text{ and } j \in B} d_{ij} \tag{50}$$

where complete linkage takes the maximum distance:

$$d_{CL}(A,B) = \max_{i \in A \text{ and } j \in B} d_{ij} \tag{51}$$

Average linkage considers the average distance between all pairs of points in clusters:

$$d_{AL}(A,B) = \frac{1}{N_A N_B} \min_{i \in A \text{ and } j \in B} d_{ij}$$

$$d_{AL}(A,B) = \frac{1}{\mid A \mid\mid B \mid} \sum_{i \in A} \sum_{j \in B} d_{ij} \tag{52}$$

where $\mid A \mid, \mid B \mid$ are the numbers of points in cluster A and B respectively.

Divisive cluster has not been explored in that extend yet. The simplest approach to apply divisive clustering could be to recursively apply the k-means method with $K = 2$ at each step [69].

B1.2.3. Density-Based Spatial Clustering of Applications with Noise

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm. Firstly, two key concepts need to be defined. Let $d(p,q)$ be the distance two points p, q. A point p is density-reachable if there exists a sequence of n points $p_1, p_2, \ldots, p_n$ and each point in the sequence is within a specified distance from the next point, thus $d(p_i, p_j) < \varepsilon$. Two points p and q are density-connected if there exists a point o such that both p and q are density-reachable from o.

Instead of defining how many clusters are needed, DBSCAN requires two hypermeters to be defined ε, n, where ε is the maximum distance between two points to be considered in the same "neighborhood", and the minimum number of points (i.e. $MinPts$), which is the minimum number of points required to form a dense region.

Having the above concepts set, a formal characterization of the points can be implemented. Core point is point if it has at least a $MinPts$ points within its ε-neighborhood. Border point a point which is within the ε-neighborhood of a core point but is not a core point. Lastly noise point is neither a core point nor a border point. The above definitions are shown in Figure 23.

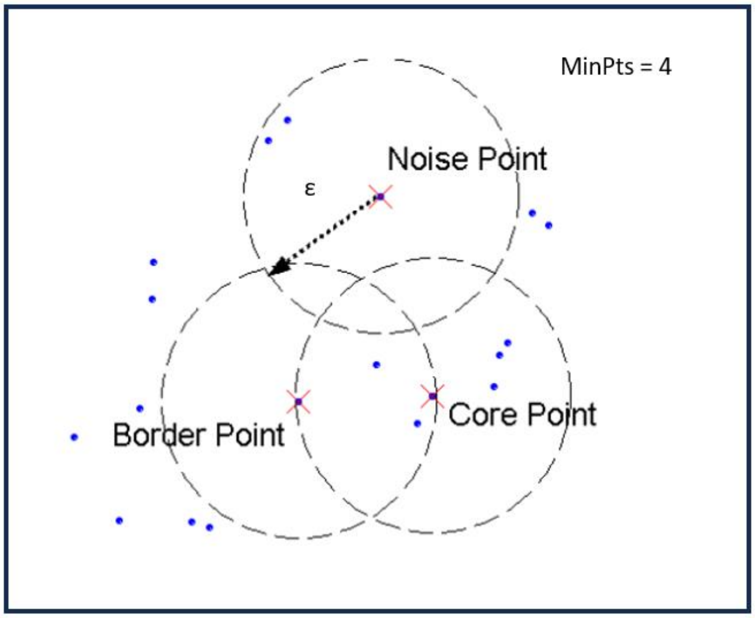


Figure 23. Definitions of “core”, “border” and “noise” points in Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, for a fixed ε when MinPts equal to 4. The core point has at least 4 points ε-close to it, the border points is exactly at the ε-border and the noise point is neither core point nor border point.

The DBSCAN pseudo-algorithm is as follows:

a) Choose randomly a point p.

b) With respect to p, retrieve all points that are density-reachable taking into account ε and $MinPts$.

c) Form a cluster if p is a core point.

d) If p is a border point, no points are density-reachable from p, and DBSCAN visits the next point of the dataset.

e) Repeat steps a-d for all points.

The determination of $MinPts$ usually comes from domain knowledge or a rule of thumb is used, where $MinPts\ = D\ +\ 1$, where D, is the dimensionality of the data. A popular method for choosing ε is the “k-distance graph”, where distance of each point to its k-th neighbor is plotted, where k is typically set as $MinPts\ -1$. A k-distance graph is shown in Figure 24:

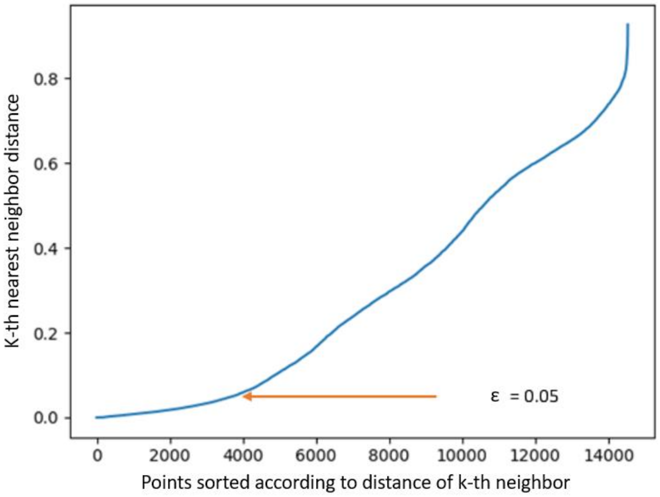


Figure 24. k-th distance graph with ε identified as the elbow point. The sorted k-th distances on the y-axis and the data points on the x-axis (sorted by their k-th distance).

B1.2.4. Principal Components Analysis and Factor Analysis of Mixed Data

Principal components analysis is a dimensionality reduction technique. Assuming that have m vectors, $x_1, x_2, \dots, x_m$ in $\mathbb{R}^d$, the goal is to apply a linear transformation and map these m vectors in space $\mathbb{R}^n$, where $n < d$, while keeping as much information as possible (i.e. maximizing the variance). The result of PCA is the principal components (PC) which are linear combinations of the input features and they are uncorrelated to each other. PCs are ordered by the amount of variability they capture from the original data.

The first PC is the linear combination:

$$Z_1 = \varphi_{11}X_1 + \varphi_{21}X_2 + \cdots + \varphi_{m1}X_m \tag{53}$$

where $\sum_{j=1}^{m} \varphi_{j1}^2 = 1$ and φ_{j1} are the "loadings" of the fist principal component and $\varphi_1 = (\varphi_{11}\ \varphi_{21} \dots \varphi_{m1})^T$ is the principal component loading vector of the 1st principal component.

The constraint $\sum_{j=1}^{m} \varphi_{j1}^2 = 1$ is necessary, since setting these elements arbitrarily large will result in arbitrarily large variance.

To compute the principal components the data need to standardized, that is subtracting the mean and dividing by the standard deviation:

$$X_{cent} = X - \bar{X} \tag{54}$$

$$X_{standardized} = \frac{X_{cent}}{\sigma_x} \tag{55}$$

And compute the covariance matrix:

$$\Sigma = \frac{1}{n-1} X_{standardized}^{T} X_{standardized} \tag{56}$$

Σ can be decomposed as follows:

$$\Sigma V = V\Lambda \tag{57}$$

where $V = (v_1\ v_2 \ldots v_n)^T$, v are the eigenvectors, and Λ is the diagonal matrix with the eigenvalues λ in its diagonal. $Z = X_{standardized}V$ corresponds to the matrix with the principal components, where $Z = (z_1 z_2 \ldots z_n)^T$.

Each principal component explains a proportion of the variance of the original data. The variance explain of the kth principal component is given by $\frac{\lambda_{\kappa}}{\sum_{i=1}^{n} \lambda_{\iota}}$. Figure 25 illustrates 2 principal components in a 3-dimensional space.

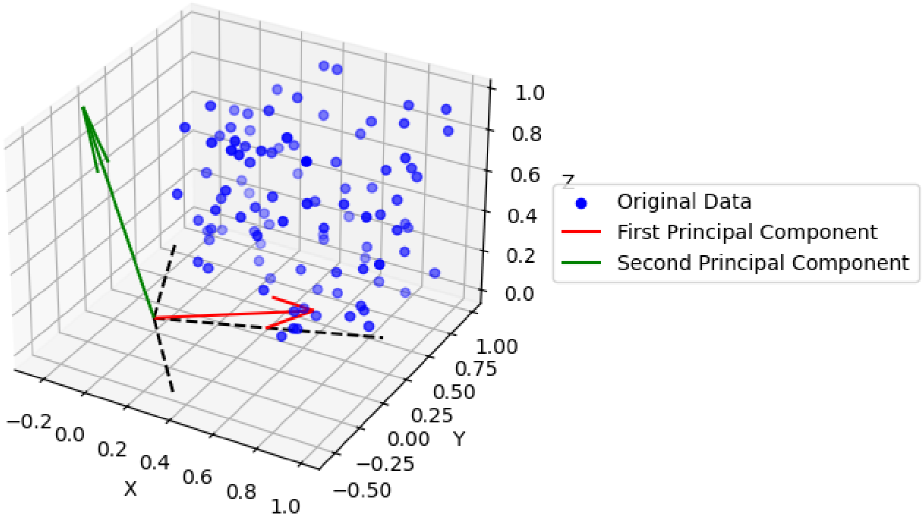


Figure 25. Two principal components in a 3-dimensional space. The two principal components are orthogonal to each other.

A graphical tool that is often used in PCA, is biplot. Using biplot the relationship between variables and observations in a reduced-dimensional space can be visualized. This is achieved by transforming using (50), the coordinates of the original datapoints into the coordinates of the reduced dimensional space (i.e. scores) and plotting them together with the loadings, as shown in Figure 26:

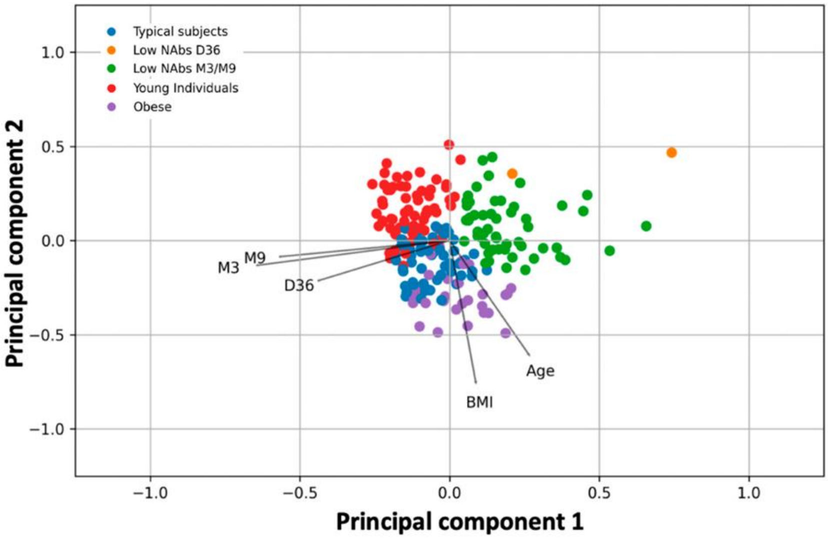


Figure 26. Biplot after applying PCA with two principal components. In the x-axis is the first principal component and in the y-axis is the second principal component. The datapoints are transformed from the original space to the reduced space and are shown as dots whereas the vectors are the loadings of each feature from the original space to the reduced space.

The number of principal components chosen, varies, depending on the use case. Often two or three principal components are chosen, because it is possible to visualize, but sometimes a scree plot used, where, the eigenvalue is plotted against the principal components, and the number of optimal principal components is decided using the elbow rule, as shown in Figure 27.

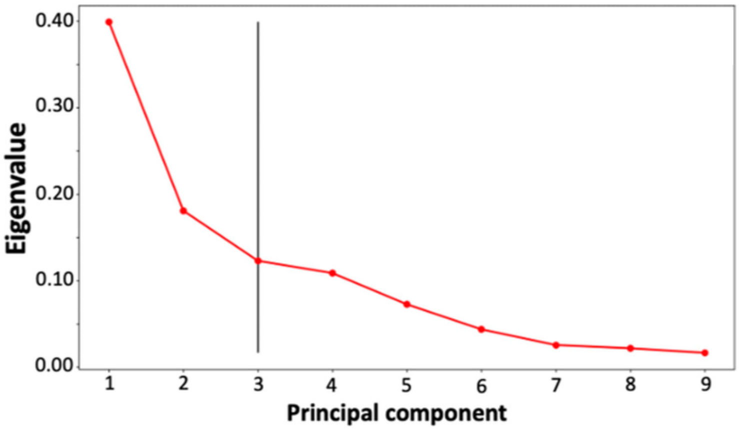


Figure 27. Scree plot example to choose the optimal number of principal components using the elbow rule. The optimal number of principal components is 3.

Even though PCA works very well in continuous data, it works poorly for mixed type of data, that is including both continuous and categorical data. Factor Analysis of Mixed Data (FAMD) is an extension of PCA, that handles mixed type of data. FAMD applies “one-hot-encoding” in the categorical data, and later combines the standardized quantitative variables (52) and the one-hot-encoded into a single matrix and performs Singular Value Decomposition (SVD):

$$X = U\Sigma V^T \tag{58}$$

where the factor scores are calculated as $U\Sigma$ and the loadings are calculated as $V\Sigma$.

B2. Deep learning

Deep learning is a specialized subset of machine learning that has garnered considerable attention with recent advancements in computational power. Its primary focus is on utilizing neural networks to model and interpret complex patterns within data. The structure and functionality of these neural networks are inspired by the human brain, enabling deep learning models to perform sophisticated pattern recognition and decision-making tasks.

B2.1. Neural networks

Neural networks are the cutting-edge models of artificial intelligence and machine learning. They consist of interconnected units called neurons, arranged in layers. There are three types of layers, input, hidden and output. The number of hidden layers can vary, depending on the use case. Each neuron in a layer connects to neurons in the next layer through weighted edges. An overview of a typical structure of a neural network is shown in Figure 28.

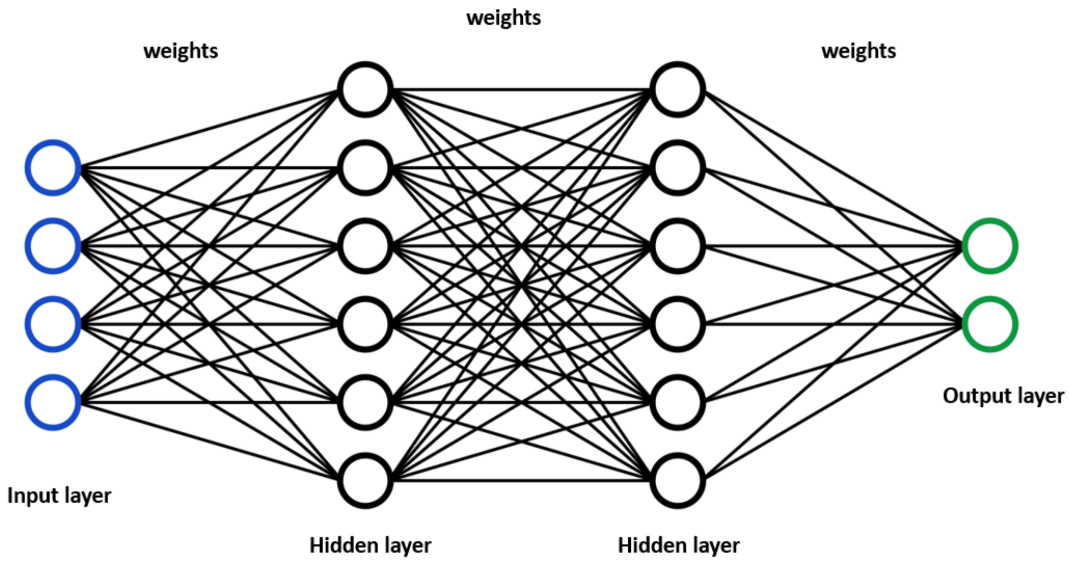


Figure 28. A typical architecture of a fully connected Neural Network. All nodes from the input layer are connected with all the nodes of the first hidden layer and all the nodes from the first hidden layer are connected with all the nodes of the second hidden layer.

The data enter the neural network from the input layer, where each neuron corresponds to a feature of the input data. The data then moves through the hidden layers, where the actual computation and learning occur. The hidden layers can vary in number and size, and each neuron in these layers applies a transformation to the input it receives. This transformation is governed by an activation function, which introduces non-linearity into the model, enabling it to learn complex patterns. Common activation functions include the sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{59}$$

Which maps input values to [0,1], which is usually used for binary classification tasks. Another common activation function is the hyperbolic tangent (tanh) function:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{60}$$

which maps the input values to [-1,1]. Lastly the rectified linear unit (ReLU) function is commonly used:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{61}$$

The model is learning through forward propagation. During this process, data enter the network through the input layer, passed through the network to finally generate an output. The input of each neuron in any layer is the output of the ones in the preceding layers after the activation function is applied. Mathematically, the output of a neuron can be expressed as:

$$a_j^{(l)} = \mathrm{f}\left(\sum_i w_{ij}^{(l-1)} a_i^{(l-1)} + b_j^{(l)}\right) \tag{62}$$

where $a_j^{(l)}$ is the activation of the j-th neuron in layer l, $w_{ij}^{(l-1)}$ is the weight of the connection between the i-th neuron in layer $l-1$ and the j-th neuron in layer l, $a_i^{(l-1)}$ is the activation of the i-th neuron in layer $l-1$, $b_j^{(l)}$ is the bias term for the j-th neuron in layer l and f is the activation function, as shown in [79]. The bias term allows the activation function to be shifted, providing additional flexibility in the learning process [79].

The learning from forward propagation, is later optimized (i.e. improved) through backward propagation. During this mechanism the model learns from its errors. Following forward propagation, the network's output is evaluated using a loss function, which quantifies the discrepancy between the forecasted and observed values. Similarly to machine learning, loss functions that are commonly used are the MSE for regression tasks:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{63}$$

where y_i is the actual value and $\widehat{y}_i$ is the predicted value and cross-entropy loss for classification tasks:

$$\text{Cross-Entropy} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\widehat{y}_i) + (1 - y_i)\log(1 - \widehat{y}_i)] \tag{64}$$

where y_i is the actual label and $\widehat{y}_i$ is the predicted probability.

Once the loss is computed, backward propagation is used to update the weights and biases in the network to minimize this loss. This is done by calculating the gradient of the loss function in reference to each weight and bias, and then adjusting the weights and biases in the opposite direction of the gradient. This process is known as gradient descent. Mathematically, calculating the gradient concerning weight $w_{ij}^{(l)}$ is given by:

$$\frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}} = \frac{\partial \mathcal{L}}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{(l)}}{\partial z_j^{(l)}} \cdot \frac{\partial z_j^{(l)}}{\partial w_{ij}^{(l)}} \tag{65}$$

where $\mathcal{L}$ is the loss function, $a_j^{(l)}$ is the activation of the j-th neuron in layer l, and $z_j^{(l)}$ is the weighted sum of inputs to the j-th neuron in layer l. The chain rule is used to compute these partial derivatives, allowing the gradients to be propagated backward through the network. This procedure is presented in Figure 29:

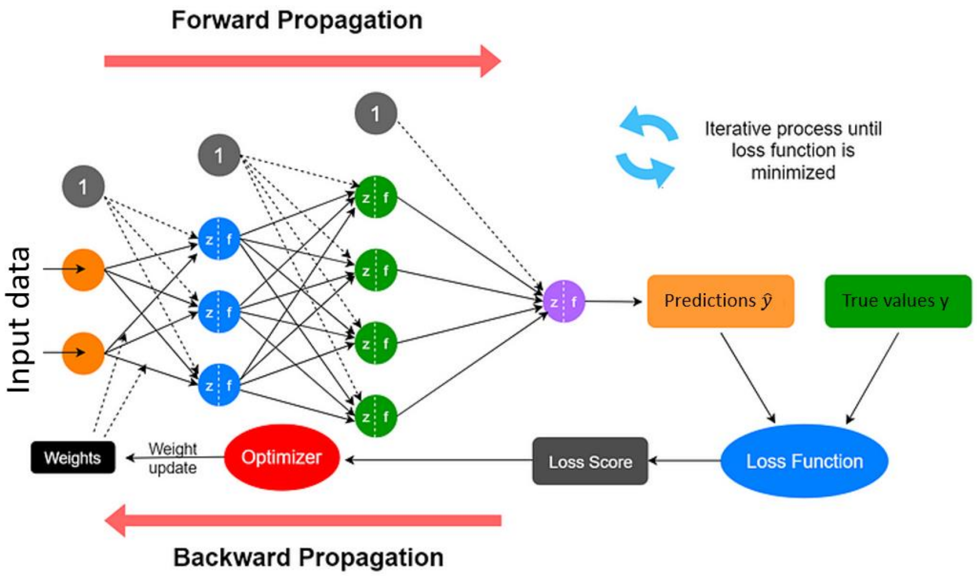


Figure 29. Illustration of forward propagation and backward propagation. During forward propagation the input data are passed through each layer of the Neural Network, weights are calculated and finally a prediction is outputted. The loss score is calculated using the prediction and the true value, with respect to the chosen loss function. Finally, the weights are updated, with the aim to reduce the loss score, during backward propagation.

The learning rate, a hyperparameter that controls the size of the weight updates, plays a crucial role in the training process. If the learning rate is too high, the network may converge too quickly to a suboptimal solution or even diverge. If the learning rate is too low, the training process may be excessively slow and may get stuck in local minima. Techniques such as learning rate schedules and adaptive learning rate methods, like Adam and RMSprop, are often used to address these issues [79].

B2.1.1. Autoencoders

A neural network with a special structure that is used in unsupervised use cases, is autoencoders. Autoencoders are designed to learn efficient representation (i.e. encoding) of the input data. They consist of two main components, the encoder and the decoder, where one is a mirrored image of the other. 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. Since the goal of autoencoders is to reconstruct the input, a common choice of loss function, $\mathcal{L}_{\text{AE}}$, is the mean squared error:

$$\mathcal{L}_{\text{AE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{66}$$

where x_i, represents the input data, $\hat{x}_i$, the reconstructed data and n the total number of observations.

The architecture of autoencoders is presented in Figure 30.

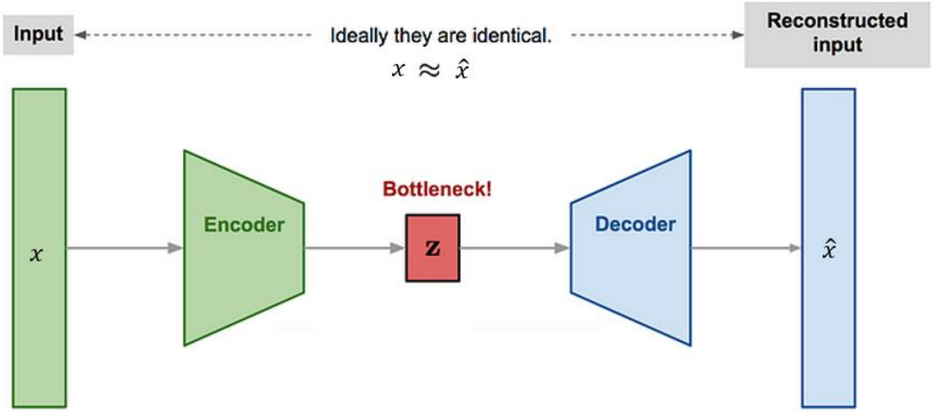


Figure 30. Visual representation of an autoencoder. The input data are compressed through the encoder while the decoder later decompresses them and tries to reproduce the input as closely as possible.

B2.2. Generative Neural Networks

A special category of neural networks is generative neural networks, which generate new, synthetic data, with the use of real ones. These models are particularly effective in fields such as image synthesis, text generation, and music composition. The main goal of generative models is to understand the underlying distribution of the training data, allowing them to generate new, similar data points.

B2.2.1. Variational Autoencoders

An extension of autoencoders, which is used for generative purposes, is variational autoencoders, which impose a probabilistic structure on the latent space. Instead of mapping input data to a single point in the latent space, VAEs map it to a distribution, typically a Gaussian distribution. This allows VAEs to generate new data samples by sampling from this distribution. Similarly to AEs, VAEs consist of an encoder and a decoder, though when minimizing the reconstruction error, a regularization terms ensures that the latent space follows a standard normal distribution, thus loss function of VAEs can be written as:

$$\mathcal{L}_{\text{VAE}} = \frac{1}{n}\sum_{i=1}^{n}\left[\text{MSE}(x_i, \hat{x}_i) + \text{KL}\big(q(z \mid x_i) \parallel p(z)\big)\right] \tag{67}$$

where KL, is the Kullback–Leibler divergence:

$$\text{KL}\big(q(z \mid x) \parallel p(z)\big) = -\frac{1}{2}\sum_{j=1}^{d}\left(1 + \log\big(\sigma_j^2\big) - \mu_j^2 - \sigma_j^2\right) \tag{68}$$

where and μ_j and σ_j are the mean and variance of the latent variable z of the j-th dimension [82]. A visual representation of VAEs is presented in Figure 31.

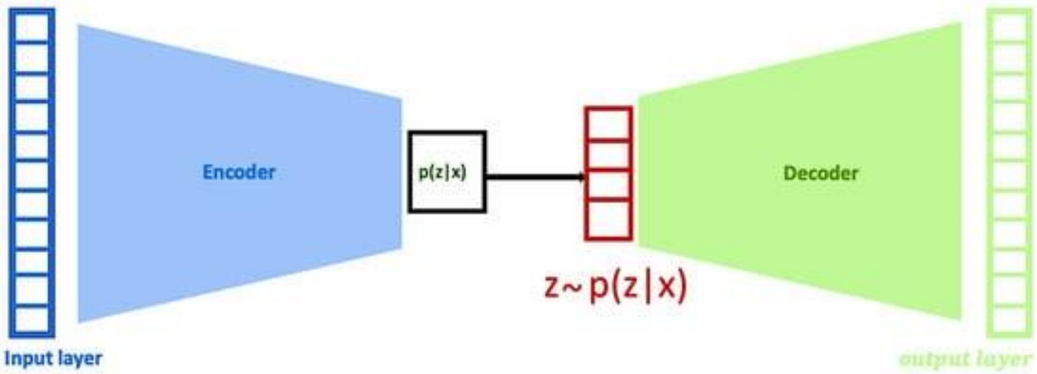


Figure 31. VAEs architecture. The data goes from the input layer to the encoder, where they are mapped to the latent space. Finally, the data are decoded from the decoder to get the output.

B2.2.2. Generative Adversarial Networks

Generative Adversarial Networks is a neural network, mainly used to generate new data that is like a given dataset. GANs are made up of two competing neural networks: a generator and a discriminator. These two networks engage in a dynamic game, where they optimize opposing objectives.

The generator, denoted as G, generates synthetic data by transforming random noise, typically sampled from a prior distribution $p_z(z)$ (such as Gaussian or uniform), into data points resembling the real data. The discriminator, represented by D, classifies input data as either real (from the actual dataset) or fake (produced by the generator). The generator's goal is to generate data, that can deceive the discriminator into classifying it as real, while the discriminator strives to correctly identify the fake data. This adversarial relationship is the key feature of GANs.

The objective of the GAN framework can be expressed as a minimax optimization problem, where G aims to minimize the probability that D correctly identifies its outputs as fake, while the D attempts to maximize its ability to differentiate real data from fake data. As shown in [83], mathematically, this can be written as:

$$\min_G \max_D V(D, G) = E_{x \sim p_{\text{data}}(x)}[\log D(x)] + E_{z \sim p_z(z)}\left[\log\left(1 - D\big(G(z)\big)\right)\right] \tag{69}$$

In Eq. 69, $p_{\text{data}}(x)$ is the distribution of the real data, and $D(x)$ is the discriminator's assessment of whether x is real. The term $G(z)$ represents the data generated by the generator from noise input z, and $D(G(z))$ is the discriminator's prediction on the generated data. The discriminator is trained to maximize $D(x)$ on real samples and minimize $D(G(z))$ on fake samples, while the generator works to maximize $D(G(z))$, effectively trying to fool the discriminator [83].

GAN training alternates between updating the discriminator and the generator. The discriminator is trained to better identify real data, and the generator is designed to create data that more convincingly imitates the real distribution. Ideally, this leads to a balance where the generator produces data so realistic that the discriminator can no longer reliably differentiate between real and generated samples.

B3. Natural Language Processing and Large Language Models

Natural Language Processing (NLP) is a specialized area within artificial intelligence that focuses on enabling computers to interact with humans using natural language. The primary aim of NLP is to allow machines to understand, interpret, and generate human language in a way that is both meaningful and useful. This involves a variety of tasks such as speech recognition, language translation, sentiment analysis, and text summarization. NLP combines computational linguistics, which models human language using rules and algorithms, with machine learning, which allows systems that leverage data and enhance over time.

One of the significant challenges in NLP is managing the ambiguity and variability inherent in human language. Words can have multiple meanings depending on the context, and the same concept can be expressed in numerous ways. To tackle these challenges, NLP systems often rely on extensive datasets and sophisticated algorithms to analyze and understand language patterns. These systems employ techniques such as tokenization, which breaks down text into individual words or phrases, and parsing, which examines the grammatical structure of a sentence.

In recent years, the development of large language models (LLMs) has significantly propelled the field of NLP forward. LLMs are AI models that used vast amounts of text data to understand and generate human language. These models utilize deep learning techniques, particularly neural networks, to learn the statistical properties of language. One of the most renowned LLMs is GPT-3, developed by OpenAI which uses 175 billion parameters, which makes it extremely powerful. It can perform a wide array of language tasks, from writing essays and poems to answering questions and generating code.

The training process for LLMs involves feeding the model with extensive text data and adjusting the model's parameters to produce predictions as close as possible to the real text. This process, known as supervised learning, enables the model to learn the patterns and structures of language. Once trained, the model can generate text that is coherent and contextually relevant. However, training LLMs requires substantial computational resources and large datasets, which can be a barrier for many organizations.

LLMs have a broad range of applications across various industries. In customer service, they can be used to develop chatbots that handle customer inquiries and provide support. In healthcare, they can assist in diagnosing diseases by analyzing medical records and literature. In finance, they can be used to analyze market trends and generate investment recommendations. LLMs can also be utilized in creative fields, such as writing and music composition, to generate new content and ideas.

Despite their impressive capabilities, LLMs also present limitations and challenges. One of the primary concerns is the potential for bias in the models. Since LLMs are trained on large datasets that may contain biased or unrepresentative data, they can inadvertently learn and propagate these biases. This can lead to unfair or discriminatory outcomes, particularly in sensitive applications such as hiring or lending. To address this issue, researchers are developing techniques to detect and mitigate bias in LLMs, such as using more diverse and representative training data and implementing fairness constraints in the training process.

Another challenge is the interpretability of LLMs. Due to their complexity and large number of parameters, it can be difficult to understand how these models make decisions and generate text. This lack of transparency can be a barrier to their adoption in certain applications, particularly those that require a high level of trust and accountability. Researchers are working on developing methods to improve the interpretability of LLMs, such as using attention mechanisms to highlight the parts of the input text that the model is focusing on.

In addition to these challenges, there are also ethical considerations related to the use of LLMs. For example, the ability of LLMs to generate realistic and coherent text raises concerns about the potential for misuse, such as generating fake news or deepfake content. There are also concerns about the environmental impact of training large models, which requires significant computational resources and energy consumption. To address these issues, researchers and policymakers are developing guidelines and regulations to ensure the responsible and ethical use of LLMs.

Chapter C: Results

C1. Machine Learning for Identifying Key Factors Influencing Neutralizing Antibody Levels After SARS-CoV-2 Vaccination

C1.1. Introduction

The novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has caused a worldwide epidemic and has become a serious global public health threat [84,85]. The coronavirus genome encodes four different major structural proteins: spike, envelope, membrane, and nucleocapsid. ACE2 receptors are typically found on epithelial cells of the oral mucosa and alveolar lung cells II but can also be found in other human organs [86]. The virus enters the body via the viral S protein and binds to ACE2 receptors [86]. COVID-19 is a systemic disease that causes both short- and long-term symptoms [87,88,89]. According to the literature, the vast majority of affected individuals have mild to moderate symptoms, with 5% to 10% having a severe or life-threatening course of disease. Worldwide, the development of effective and safe vaccines and drugs, as well as new diagnostics and therapies, is a top priority.

Vaccination against SARS-CoV-2 with BNT162b2 mRNA vaccine is playing a critical role in most countries [90,91,92]. After immunization, healthy individuals exhibit significant levels of IgG antibodies and neutralizing antibodies directed against the spike receptor binding domain of SARS-CoV-2, as well as a protracted B-cell response in the germinal center [93,94]. It is worth noting that NAbs content has been associated with clinically significant immune protection against COVID-19 [95,96].

Although BNT162b2 is highly effective against COVID-19, a time-dependent decrease in antibody levels against SARS-CoV-2 was observed in vaccinated individuals [97-101]. Even one month after the second BNT162b2 injection, there was a modest decline in antibody titers, and the time since the second vaccine dose was associated with decreased neutralizing antibody activity against SARS-CoV-2 variants and attenuated protection against COVID-19 [97-101]. Although new SARS-CoV-2 variants threaten the high level of immune protection achieved through vaccination, [99, 102] a booster vaccine dose combined with transmission-reducing behaviors remains effective in preventing COVID-19 [103, 104]. Moreover, the persistence of the cellular response to SARS-CoV-2 after vaccination could influence both the humoral response and the overall protection against COVID-19 [101, 105-107].

This study had two goals. The first one was to investigate the kinetics of NAbs and anti-S-RBD IgGs against SARS-CoV-2 after full vaccination with the BNT162b2 mRNA vaccine for up to 9 months in healthy individuals. The second goal was the identification of characteristics of individuals that may act as predictive factors for neutralizing antibody levels long after vaccination.

A useful tool to answer the latter is the application of machine learning techniques, a subfield of artificial intelligence [108,109,110]. The goal of ML is generally to understand the structure of data and to fit that data into models that can be understood and used by humans. ML techniques allow data to be trained and then statistical analysis applied to produce values that fall within a certain range [109,110]. Thus, ML enables the development of models from sample data to automate decision-making processes based on data inputs. Based on how the system learns or receives feedback on what it learns, ML tasks are generally classified into broad groups. The two most commonly used ML methods are "supervised learning," in which algorithms are trained based on labeled input and output data, and "unsupervised learning," in which the algorithm is not provided labeled data to discover structure in the input data [109,110].

To identify factors responsible for NAbs levels nine months after vaccination with BNT162b2, four popular ML methods were applied on the data from the to unveil hidden relationships between subject characteristics and NAbs that are difficult or impossible to find using classical statistical approaches. To take advantage of both supervised and unsupervised methods, a combination of both methods was used in this analysis. In addition, three unsupervised methods were used, which complement each other.

C1.2. Materials and Methods

C1.2.1. Data acquirement and description

The data used in this study were leveraged from a clinical study that was designed to determine the kinetics of anti-SARS-CoV-2 antibodies after COVID-19 immunization with the BNT162b2 mRNA vaccine (NCT04743388). The ethics committee of the General Hospital Alexandra approved the study protocol (Ref No. 15/23 December 2020). The study was carried out in accordance with the Declaration of Helsinki and the International Conference on Harmonization for Good Clinical Practice standards of care. All participants provided written informed consent at study entry.

The dataset included 309 patients' demographic information, medical history, and prescriptions were collected during an interview at study entry. Weight and height were used to compute the individual's body mass index (BMI). All of the individuals were assigned to 1 of 3 groups based on their BMI: BMI ranges from 18.5 to 24.9 for those who are underweight; 25 to 29.9 for those who are overweight; and 30 or more for those who are obese. The afore mentioned characteristics of the dataset are displayed in Table 1:

Table 1. Characteristics of the Participants in the Study.

Participant Characteristic	Value*
Sample size	309
Gender	
Men	107 (34.6%)
Women	202 (65.4%)
Age (median)	48.0
Body mass index (median)	24.8
Underweight (n, %)	15 (4.85%)
Normal weight (n, %)	148 (47.90%)
Overweight (n, %)	102 (33.01%)
Obese (n, %)	44 (14.24%)

***Values in parentheses refer to percentages**

C1.2.2. Population modeling and simulations

Individual longitudinal percent inhibition values were explored in terms of population kinetic analysis, utilizing the stochastic approximation expectation maximization methodology for nonlinear mixed effects, followed by importance sampling approaches. Because the goal of this study was to determine the pace at which NAbs was eliminated from the body, only the dropping portion of the NAbs levels was modeled. As a result, data from 2 weeks after the second immunization up to 9 months after the second vaccine were used.

A number of structural models, including 1- and 2-compartment designs, were evaluated. Single exponential or linear functions, as well as piecewise linear functions, were used to represent the kinetics of NAbs elimination. The NAbs levels were classified as normal or log-normal, and a variety of residual error models were examined (e.g., constant, proportional, and combined). Following the construction of the final optimal structural model, the impact of the individuals' characteristics (e.g., age, gender, BMI) on the model parameters was investigated. The Wald test was performed to examine whether or not variables could be utilized to explain variation in the parameters. This was accomplished entirely by writing the appropriate code in Monolix 2020R1 Mlxtran language (Lixoft, Orsay, France). To predict NAbs concentration levels at 12, 15, and 18 months after vaccination, simulations were conducted assuming that the kinetic parameters remained unaltered after the ninth month following the construction of the population model. A number of 1000 individuals were simulated using resampling from the same individual pool with replacement. A sample size of 1000 individuals were used for the simulations, and the individual model parameter values were randomly drawn to allow for robust prediction. However, since the simulated sample size is much larger than the original one, the individual parameter values were drawn from the original set with replacement, that is, each subject can be used more than once. The simulations were carried out using Simulx (Lixoft, Orsay, France).

C1.2.3. Machine Learning

ML algorithms were used to analyze the percent inhibition thresholds (i.e., NAbs) with respect to all data from the above subjects. In ML, tasks are divided into two categories: supervised learning and unsupervised learning. In supervised learning, the computer is fed examples of inputs labeled with expected outputs. The goal of this method is for the algorithm to "learn" by comparing its real outputs to the "learned" outputs to detect errors and adjust the model accordingly. The supervised learning algorithm used in this study was Random Forest.

Since the data is not labeled in unsupervised learning, it is left to the learning algorithm to discover commonalities between the input data. Unsupervised learning can have the simple purpose of detecting hidden patterns in a data set, but it can also have the goal of feature learning, which allows the computational engine to automatically discover the representations needed to classify the raw data. In this work, k-means cluster analysis and two-dimension reduction approaches (principal component analysis and factor analysis of mixed data) were used.

Before applying the ML approaches, the data were tested for multi-collinearity, i.e., the characteristics of the independent participants were examined to see if they correlated with each other. Multi-collinearity has a significant impact on the variance associated with the problem and may also affect model interpretation by undermining the statistical significance of the independent variables. For all pairs of variables, Spearman's correlation coefficient (i.e., r) was used to estimate correlation.

C1.2.4. Principal Component Analysis

Principal component analysis is a popular approach to transform a high-dimensional set of features into a low-dimensional set of features [62, 111]. The goal of PCA is to find the lowest dimensional representation of the data while capturing as much information/variance as possible.

PCA transforms the original space formed from the original dataset into a new space that is a linear combination of the dimensions of the dataset. Each additional dimension that is developed is called a principal component. The new coordinates of the data are called "scores".

PCA is a technique for reducing the dimensions of a data matrix in an attempt to capture as much variability as possible. Each PC explains a portion of the variance in the original data set. The direction of the first principal component is the direction in which the data varies the most. The contribution of each original dimension to the new dimension is quantified by various factors called "loadings." Each principal component is the normalized linear combination of the original features, where normalized means that the squared sum of the loadings of each principal component equals one. The closer the loading value is to +1 (or 1), the more positive (or negative) the contribution of that feature to PC.

The best way to visualize the loadings and scores together is to use the "biplot". The biplot is a two-dimensional scatter plot where the two axes represent the two most important PCs in terms of variance explained. The data points are plotted in this two-dimensional coordinate system, using the scores as coordinates, and the loadings of the first two PCs of each trait are plotted over the data points.

Scree plots were used to determine the least number of principal components needed to accurately represent the original data. The main purpose of a scree plot is to show the results of the component analysis and to locate the apparent change in slope (elbow). In a scree plot, the eigenvalue is plotted against the principal components. The eigenvalue of a component divided by the sum of the eigenvalues is the proportion of the variance explained by that component. The first component usually explains a large portion of the variability, the next components explain a moderate portion, and the last components explain only a small portion of the total variability.

C1.2.5. Factor Analysis of Mixed Data

While PCA works well on continuous data, it is ineffective on categorical data because the resulting PCs aim to maximize the variability of the underlying data set in the new, transformed, lower-dimensional space. This is because categorical features require "one-hot coding" that transforms them into binary features. The concept of variability is consistent with binary features, as is the use of PCA.

Factor analysis of mixed data is a technique that combines PCA with multiple correspondence analysis to analyze numerical and categorical variables [111]. By considering both continuous and categorical data, FAMD also leads to a low-dimensional space. The results can be summarized in a biplot as in PCA, as mentioned above. In addition, as with PCA, scree plots were created to determine the required number of principal components.

C1.2.6. K-Means Cluster Analysis

When the data is unlabeled, another popular unsupervised learning technique is k-means cluster analysis, which creates groups (clusters) of variables from the original dataset [111]. K-means partitions a p-dimensional space into k groups (where p is the number of variables in the dataset). Each of the k clusters is defined by a centroid, which, as the name implies, is located at the center of the cluster. Each point in the dataset is assigned to the cluster with the closest centroid. A meaningful interpretation of the clusters is possible by looking at the coordinates of the centroid. To determine the optimal number of clusters, the "elbow" approach (scree plot) was used. For different values of k, the sum of squared distances between each point and the centroid in a cluster (Within-Cluster Sum of Squares, WCSS) is determined. A line plot is then created with the squared distances on the y-axis and the different k values on the x-axis. The ideal k is the point at which the line plot forms an "elbow" (i.e., an angle).

In addition, the silhouette score and Davies-Bouldin index were calculated to determine the optimal number of clusters. The silhouette score is used to evaluate the quality of clusters created using clustering methods, i.e., to assess how well samples cluster with other samples that are similar to each other. The silhouette score is created for each sample from each cluster and ranges from −1 to +1, with a high number (close to 1) indicating that the object matches its own cluster well. A value less than 0 indicates that the data from the clusters may not be correct. Negative values often indicate that a sample has been assigned to the wrong cluster. Similarly, the Davies-Bouldin index is a validation metric used to determine the best number of clusters. The minimum value is zero, with lower values indicating better clustering.

C1.2.7. Random Forest

In the context of machine learning, bagging is a technique in which numerous "copies" of the training data are made (each "copy" being slightly different from the others). Then a weak learner, such as a "decision tree", is applied to each copy. In this way, many weak models are generated, which are then integrated. Random forest is a bagging approach of supervised learning, where a large number of decorrelated trees are created and then combined by averaging to obtain a more accurate and stable prediction of the target variable [111]. To make the model more robust, it is common to divide the original dataset into two sections, “train” and “test”. The training dataset is used to train the model and the test dataset is used to evaluate the performance of the model. In a classification problem (i.e., when the target variable is categorical), the confusion matrix can be used to evaluate the performance of a model. The confusion matrix is an M × M matrix, where M is the number of target classes of the target variable. The matrix contrasts the true and expected classes. This provides a comprehensive view of the overall performance of the categorization model and the types of errors it makes. The accuracy and misclassification of the predictions can be calculated to reflect the performance of the classification. By examining the feature importance, it is also possible to see the contribution of each feature to the prediction of the target variable when using random forest.

C1.3. Results

C1.3.1. Modeling the kinetics of neutralizing antibodies

To describe the elimination kinetics of NAbs, a population kinetic model was developed. The final best model obtained from the population study comprised a piecewise function for the elimination constant, as well as a proportional error model (Table 2).

Table 2. Parameter Estimates for the Final Best Model Describing the Kinetics of NAbs.

Parameter	Value	Standard Error	% Relative Standard Error
Fixed effects			
No	96.05	1.690	1.76
kel1	0.017	0.001	8.48
beta_kel1_Age*	0.020	0.002	9.27
kel2	0.111	0.008	7.63
kel3	0.071	0.007	9.16
Standard deviation of random effects			
omega_No	0.016	0.002	12.8
omega_kel1	1.37	0.099	7.25
omega_kel2	0.98	0.067	6.86
omega_kel3	1.10	0.081	7.36
Correlations			
corr_kel2_kel1	0.78	0.037	4.68
Error model parameters			
b	0.043	0.001	2.79

***Indicates a statistically significant (P = 0.0037 < 0.05) contribution of “age” as a covariate to the first-phase elimination rate constant.**

b = the proportional error term of the model residual variability; beta_kel1_Age = constant for the contribution of "age" on kel1; corr_kel2_kel1 = the correlation coefficient between kel1 and kel2; kel1 = elimination rate constant of the first trimester (zero to third month); kel2 = elimination rate constant of the second trimester (third to sixth month); kel3 = elimination rate constant of the third trimester (sixth to ninth month); NAbs = neutralizing antibodies; N0 = average initial NAbs value; omega = between-subject variability estimate for each parameter.

The first elimination constant (kel1) was 0.016 NAbs/day (standard error = 0.0014) and corresponded to the early decay period lasting up to 3 months. Following that, the elimination constant was calculated to be kel2 = 0.099 NAbs/day (standard error = 0.0076). This result shows that the NAbs disappear relatively slow at first, but that their removal becomes around 6 times greater from the third to the sixth month, indicating that they are eliminated much more quickly. In the period between the sixth and ninth month, the elimination rate constant is 0.071 NAbs/day (standard error = 0.0065), which is almost 4.5 times higher than in the first trimester and 30% less than in the third to sixth trimester. This finding indicates a decrease in the elimination capacity of NAbs, which is a desirable feature, as individuals retain NAbs for a longer period.

According to the findings, age has a significant impact only on the elimination ability of the first trimester (beta kel1 = 0.019, P = 0.0037) implying that age does not contribute significantly in the elimination of NAbs after the third month. This relationship, on the other hand, was statistically significant (P < 0.001) only during the first phase of the study (up to 3 months). The mathematical model describing the influence of age on the elimination constant is as following: log(kel1) = log(0.016) + eta_kel1, where eta_kel1 refers to the random effect for intersubject variability of the elimination constant. Other variables such as gender, BMI were not shown to have a statistically significant impact on antibody kinetics. Between kel1 and kel2, it was discovered that there was a positive association (coefficient = 0.78). No correlation was found between the third elimination constant and any of kel1 and kel2.

The goodness of fit and validation plots, of the final model, are depicted in Figure 32. Figure 32A shows 3 representative individual profiles of NAbs values (% inhibition) versus time. The solid lines indicate the model's anticipated values, whereas the circles show the observed NAbs levels. The close proximity of the model's predicted line to the actual points demonstrates the model's high descriptive ability. In Figure 32B, the individual observed versus predicted NAbs values are depicted for the 3 age groups, namely, 20 to 40, 40 to 55, and ≥55-year-old category. The nice predictive performance can be validated by the fact that the vast majority of values are lying within the 90% prediction interval (dashed lines in Figure 32B).

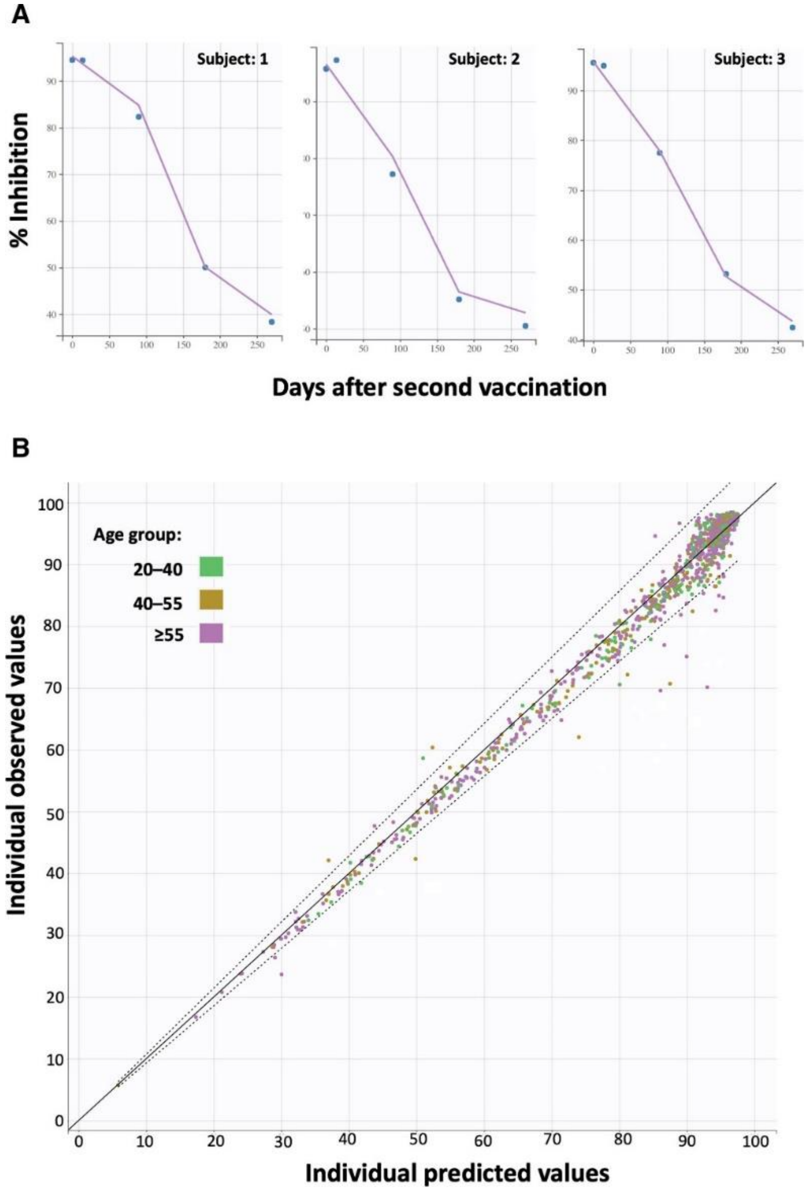


Figure 32. Indicative individual profiles of neutralizing antibodies (% inhibition) over time (A) and predicted vs observed (B). In plot A, the solid lines refer to the values predicted by the model, while the circles show the experimental values. The close transition of the model predicted line to the actual points shows the good predictive ability of the model. Plot B shows the overall fitting results for the 3 age groups, which are indicated as points with different colors. The solid line represents the optimal prediction performance, while the dotted line represents the 90% prediction interval.

C1.3.2. Simulations and predictions

Aiming to predict NAbs levels at 12, 15, and 18 months following the second vaccination, simulations were carried out using the final population model, assuming that the removal of NAbs would be the same as that found in the last trimester (i.e., from sixth to ninth month) after the second vaccination. The simulation results are depicted in Figure 33.

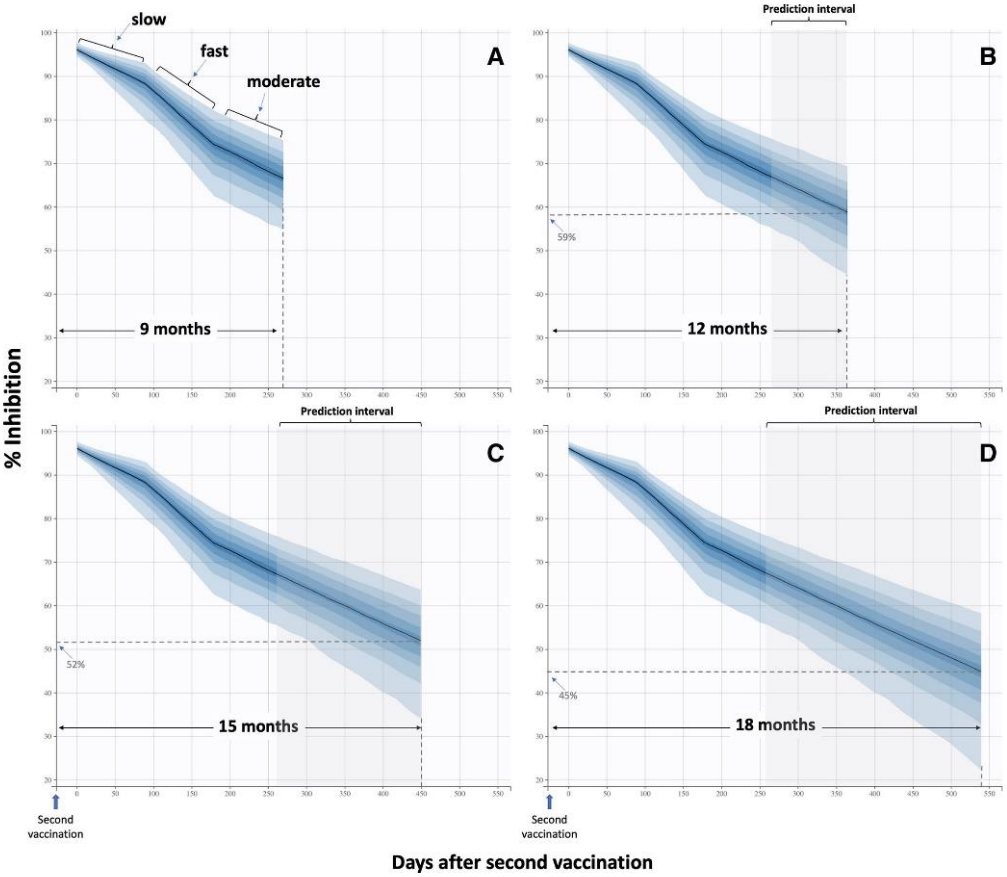


Figure 33. Simulated neutralizing antibody levels over a period of 9 months (A), 12 months (B), 15 months (C), and 18 months (D). During the first trimester, a relatively slow elimination of antibodies occurs, whereas the antibody elimination tends to increase during the next trimester (third to sixth month). In the last trimester of the study (sixth to ninth month), the elimination ability slows down by 30% compared with the second phase. Predictions for 12, 15, and 18 months assumed a similar elimination ability as the one observed in the last trimester.

Figure 33A shows the simulation prediction for 9 months for which we already have experimental data. The predicted average value of NAbs is 67.1%, which is quite close to the observed value of 65.7% (Figure 32). This result confirms the predictive ability of the model, at least for time points close to 9 months. To address the current fundamental global question of whether and when a third dose should be administered, our simulations were extended to 12, 15, and 18 months. Presumably, the accuracy of the predictions might decrease with increasing duration, as currently there is no information about the elimination ability of NAbs from the human body in long term. Nevertheless, it was worthwhile to make the predictions for longer time periods to gain insight into what we can expect in the future. In this context, Figure 33B predicts that the average percent inhibition of NAbs would be 59% 1 year after the second vaccination. This value will further decrease to 52% after 15 months (Figure 33C) and reach a value of 45% 18 months after the second vaccination (Figure 33D). These results suggest that at 1 year after full vaccination with BNT162b2, a large portion of people (around 41%) would not be highly protected. At 15 and 18 months, the risk would be even higher, as the percentage of individuals who are not highly protected might be as high as 48% and 55%, respectively.

C1.3.3. Neutralizing Antibody Levels

Figure 34 shows the mean NAbs values (±standard deviation) from the day of the first vaccination to nine months after the second vaccination with the BNT162b2 mRNA vaccine.

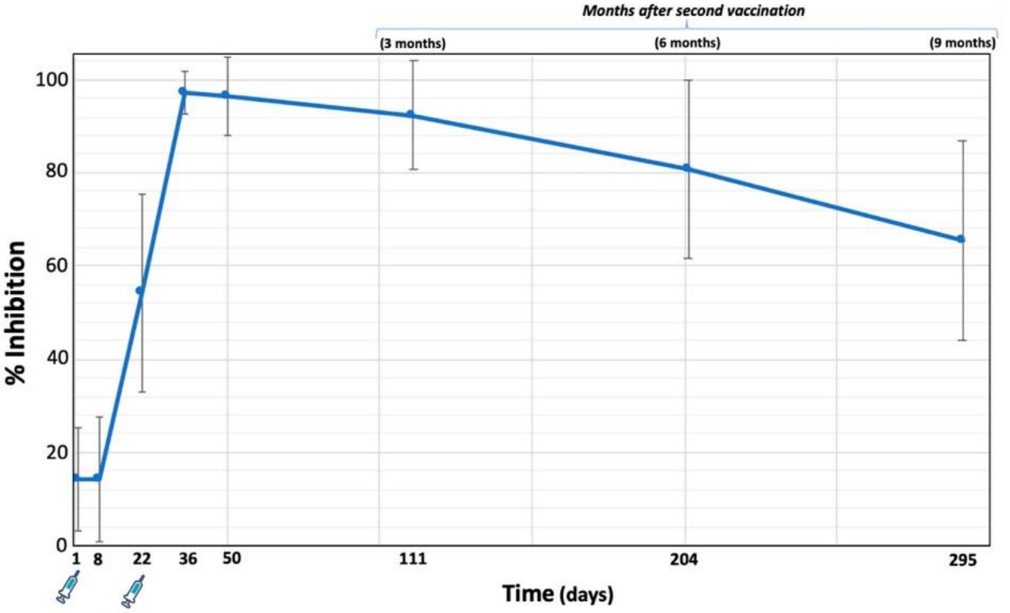


Figure 34. Mean percent inhibition (±standard deviation) of SARS-CoV-2 binding to the human host receptor angiotensin converting enzyme-2 after vaccination with the BNT162b2 mRNA vaccine in 302 subjects. Neutralizing antibody levels were measured on day 1 (first vaccination day), 8, 22 (second vaccination day), two weeks later, and one month, three months, six months, and nine months (i.e., 295 days from the initiation of the study) after the second vaccination.

At D1 and D8, median NAbs titers were low (14.23% and 14.28%, respectively), indicating that neutralizing antibodies are not formed within the first week after vaccination. Three weeks later, i.e., on the day of the second vaccination, the median value of inhibition was 54.19%, while two weeks later the maximum value is reached (97.24%). From this point on, the inhibition decreases continuously and nine months after the second vaccination the median value of NAbs is 65.43%.

C1.3.4. Principal Component Analysis and K-Means Clustering

To extract the participants' information and examine its relationship with NAbs levels, principal component analysis was performed. Figure 35 shows the results of the PCA analysis superimposed on those of the k-means cluster analysis. The observations (study participants) are shown as dots in the plane formed by the two principal components, whereas the lines represent the vectors of the variables, namely the NAbs levels at D36, M3, and M9, and BMI and age. The different color of the points refers to the grouping from the k-means cluster analysis. Overlaying PCA and k-means clustering plots allows simultaneous evaluation and comparison of the results from the two methods. Scree plots were created to select the optimal number of principal components (Figure A1A) and clusters (Figure A1B). Separate plots of PCA and k-means clustering can be found in Figure A2 and Figure A3 in the Appendix A.

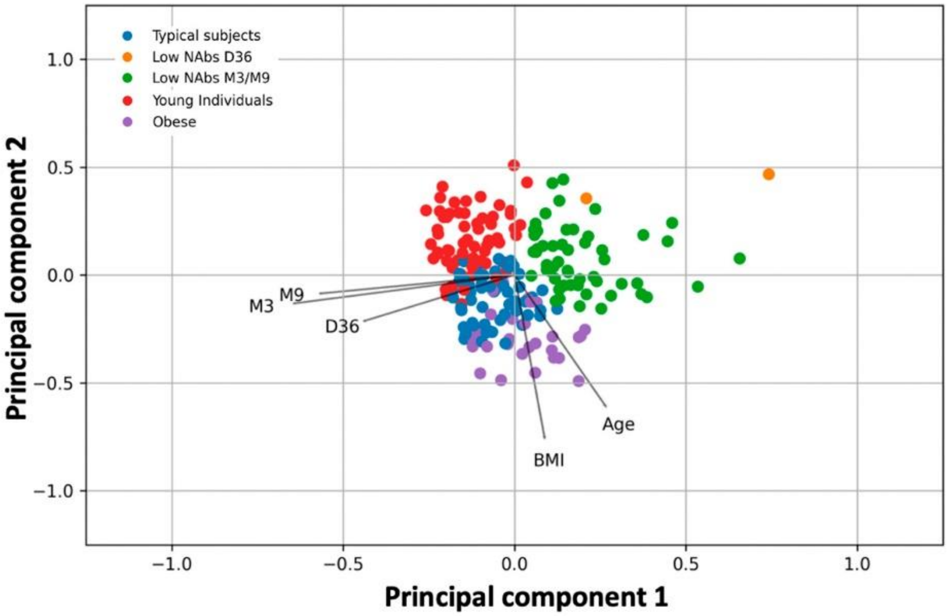
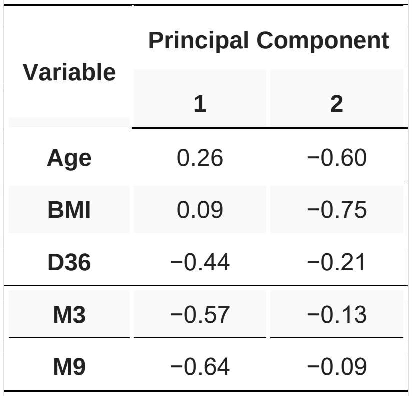


Figure 35. Principal component analysis of the features used in the study. The color of the dots is from the k-means cluster analysis, which divides the subjects into five groups, each with one distinguishing feature. The coordinates of the two-dimensional plot refer to the "scores" of the variables in the dimensionality-reduced space (two dimensions (principal components) are shown in the plot). Key: D36, neutralizing antibody levels two weeks after second vaccination; M3 neutralizing antibody levels three months after second vaccination; M9, neutralizing antibody levels nine months after second vaccination; BMI, body mass index.

The first two principal components explain 63.4% of the total variability (39.8% and 23.6% for the first and second components, respectively). Visual inspection of Figure 35 shows that M3 and M9 are adjacent to each other on the left side of the plot near the first principal component, while D36 is also close to them, indicating their strong relationship. With respect to the first principal component, the loadings of D36, M3, and M9 are −0.44, −0.57, and −0.64, respectively (Table 3). This means that someone who has high NAbs on M3 will also have high levels at M9. Similarly, high levels at D36, i.e., at the highest inhibition reached two weeks after the second vaccination, are likely to result in elevated levels nine months later.

Table 3. Loadings for the two main principal components of the final PCA model.



Variable	Principal Component	
	1	2
Age	0.26	−0.60
BMI	0.09	−0.75
D36	−0.44	−0.21
M3	−0.57	−0.13
M9	−0.64	−0.09

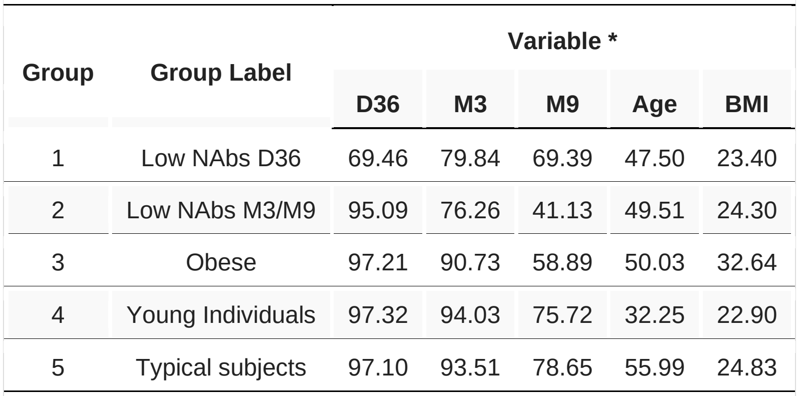
On the other hand, BMI and age are in the right/lower part of the graph with positive values with respect to the first principal component (loadings with respect to the first component: 0.26 (age) and 0.09 (BMI)). This means that BMI and age contribute negatively to the three NAbs levels. However, the angle between these two characteristics (age or BMI) and the three NAbs vectors is close to 90°, which means that their influence should be small. As expected, the variables BMI and age were found to be related, implying that as age increases, BMI also increases.

Correlation between all variables was assessed to exclude variables with strong linear relationships. Spearman's correlation coefficients (r) are shown in Figure A4. From this analysis, NAbs values at M1 (i.e., one month after the second vaccination), M3, M6, and M9 are moderately or strongly correlated. Also, NAbs values at D1 and D8 are strongly associated with each other (Spearman's r = 0.82). Therefore, the NAbs values at D36, M3, and M9 were included in the final PCA model.

In the same plot with PCA, the k-means cluster analysis yielded five natural groups of individuals according to their characteristics (age, BMI, sex, height, medical history, medication, etc.), with each group having a different color (Figure 35). These k-means cluster groups are consistent with the groups identified by PCA. The estimated silhouette score was 0.45, whereas the Davies-Bouldin score for the final best model was 0.57; both indices indicate adequate cluster discrimination. Several seeds were used for k-means clustering, all of which yielded nearly identical results, confirming the clustering results.

Within each group, the characteristics of some individuals differ in the sense that one characteristic (e.g., BMI) is different in a particular group than in the other clusters. In this context, the five natural groups refer to obese individuals, young individuals, individuals with typical characteristics (i.e., not elderly or obese, etc.), and participants with low NAbs at D6 or M3/M9 (Table 4).

Table 4. Characteristics of each group formed by k-means cluster analysis. Values refer to the mean estimates for each category.



Group	Group Label	Variable *				
		D36	M3	M9	Age	BMI
1	Low NAbs D36	69.46	79.84	69.39	47.50	23.40
2	Low NAbs M3/M9	95.09	76.26	41.13	49.51	24.30
3	Obese	97.21	90.73	58.89	50.03	32.64
4	Young Individuals	97.32	94.03	75.72	32.25	22.90
5	Typical subjects	97.10	93.51	78.65	55.99	24.83

In the first group, the distinguishing feature is mainly the low number of NAbs at D36 (mean = 69.46%) and secondly the low inhibition threshold at M3 (mean 79.84%). Of course, there is another group of subjects whose salient feature is the relatively low number of NAbs at M3 (76.26%) and M9 (mean = 41.13%). The third group of participants consists of obese subjects (mean BMI = 32.64). Young individuals with a mean age of 32.25 years form the fourth group, while the remaining participants with typical values for all five characteristics form the last group. In addition, the topological properties of the clusters and PCA vectors can be used to determine the relationship between them (Figure 35). For example, obese subjects tend to have low NAbs, whereas young subjects or subjects with typical traits have higher inhibitory values. Subjects with "typical" traits have intermediate performance, as they fall between obese and young subjects. The highest immunological response nine months after vaccination is seen in subjects belonging to the last group, namely with BMI = 24.83 and age = 55.99 years. The lowest NAbs values at M9 (and M3) are observed in subjects who have a combination of conditions: Age between 47–50 years and BMI close to 24 (i.e., groups 1 and 2 in Table 5).

C1.3.5. Factor Analysis of Mixed Data

Principal component analysis is useful for reducing the dimensionality of the data set and identifying patterns in the data. However, information from categorical variables such as sex, concomitant diseases, medications, and whether the subject has had a previous COVID19 cannot be used because PCA works only for numeric data. In addition, to investigate the possible influence of both numeric and categorical variables, a factor analysis of mixed data was performed in this study. In this case, all available information from the individuals participating in the study was analyzed. Factor analysis of mixed data was applied several times by including or excluding certain variables to find the model that explained the most variability in the data.

The biplot, score, and loading plot of the final FAMD model are shown in Figure 36, while the scree plot, constructed to select the optimal number of principal components, is shown in Figure A1C. The total percentage of variability explained by the first two principal components is 68.2% (36.4% and 31.8% for the first and second components, respectively). Visual inspection of the FAMD results shows that the common aspects are consistent with PCA. Namely, age and BMI contribute negatively to NAbs values at D36 and M9. Again, the orthogonal arrangement of the vectors BMI and D36/M9 shows the small contribution of BMI to NAbs values. The angle between the "age" vector and the two NAbs vectors is obtuse, indicating that "age" certainly makes a negative contribution to neutralizing antibodies. Young people tend to have higher neutralizing antibody levels than older people. The fact that the angle (age-0-D36) is larger than (age-0-M9) means that age affects NAbs levels in D36 more than those in M9 (Figure 36). In other words, the influence of age is stronger at time points close to the vaccination day, and the role of age seems to decrease with time.

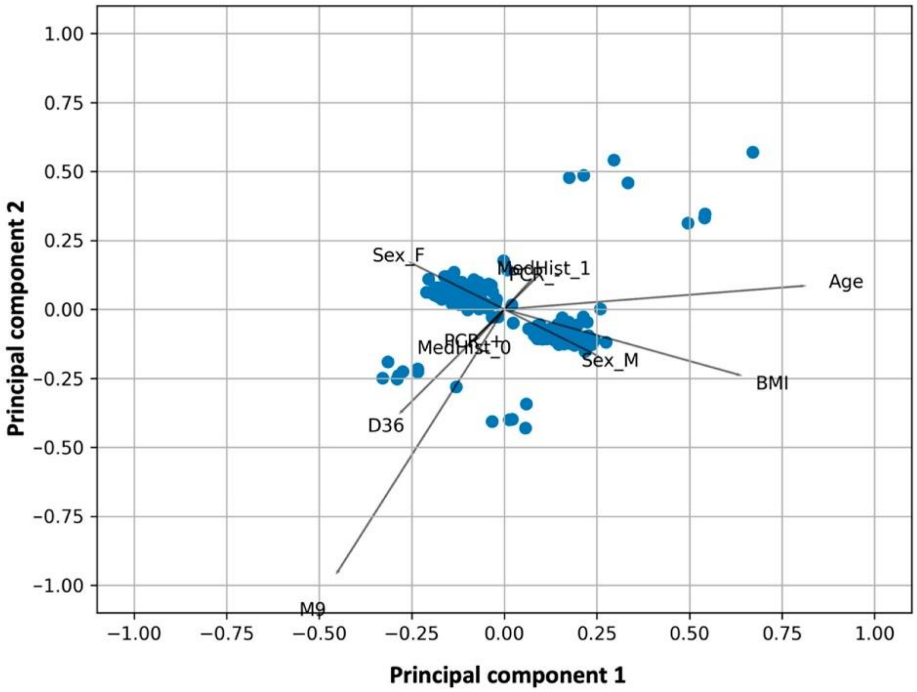


Figure 36. Biplot for factor analysis of mixed data. The coordinates of the two-dimensional plot refer to the "scores" of the variables in the dimensionality-reduced space (two principal components are shown in the plot). The lines represent the vectors of the variables in the 2D space. Key: D36, neutralizing antibody levels two weeks after second vaccination; M9, neutralizing antibody levels nine months after second vaccination; BMI, body mass index; MedHist_1, subjects with autoimmune disorders; MedHist_0, subjects without autoimmune disorders; Sex_M, men; Sex_F, women; PCR+, subjects with previous COVID-19 infection; PCR-, subjects without previous COVID19 infection.

In addition, the FAMD biplot can provide further insight into the role of categorical variables. In this context, gender does not seem to affect NAbs levels at M9 (mainly), but there seems to be a weak relationship at D36. In the latter case, it appears that female subjects (sex = 2 in the graph) are associated with higher neutralizing antibody levels two weeks after the second vaccination, i.e., at D36. In the FAMD graph, it is noticeable that the performance of the two sexes is exactly opposite because their vectors are in opposite directions, i.e., their loadings are opposite with respect to both axes.

Figure 36 also shows that a previous positive PCR test (i.e., individuals infected with SARS-CoV2) has no effect on titers at M9, but PCR+ is strongly related to inhibition values at D36. As expected, earlier infection with COVID-19 leads to higher values at D36 (mainly) and nine months later (i.e., at M9).

Finally, an important finding regarding the role of comorbidities can be seen in Figure 36. Due to the limited sample size, subjects were divided into two categories: those with autoimmune diseases and those without autoimmune diseases, i.e., those who were completely healthy or suffered from other diseases such as cardiovascular disease, mental illness, etc. FAMD analysis revealed that the vector of patients with autoimmune diseases (MedHist_1) is arranged in the opposite direction to vectors D36 and M9, indicating that subjects with autoimmune diseases have lower inhibitory levels compared to those without autoimmune problems. This feature is evident in NAbs levels early after vaccination, e.g., two weeks after the second vaccination, but to a lesser extent in NAbs titers nine months after full vaccination.

C1.3.6. Random Forest

To quantify the predictive factors for NAbs levels nine months after complete vaccination with BNT162b2, the random forest technique was used. Figure 37 shows the importance of the variables included in the analysis, ordered from highest to lowest contribution to M9 values. It is plausible that NAbs levels at M3 have the highest contribution (score = 34.2%), followed by inhibition titers at D36 (score = 22.5%). BMI and age, which already had a negative effect on M9 levels in the PCA and FAMD analyzes, also contribute to a lesser but important extent: 18.3% and 17.6% for BMI and age, respectively. Gender has a small effect on M9 values, as significance estimates were 2.0% for men and 2.0% for women. Medical history, i.e., whether you have an autoimmune or other disease (or are healthy), has a small effect on M9 levels (value = 1.0%). Finally, previous infection with SARS-CoV-2 before the first vaccination has a negligible effect on M9 inhibition titers. It should be mentioned that a 2:1 split (training set: test set) was used. In addition, multiple replicates of the 2:1 split were used to ensure the validity of the results.

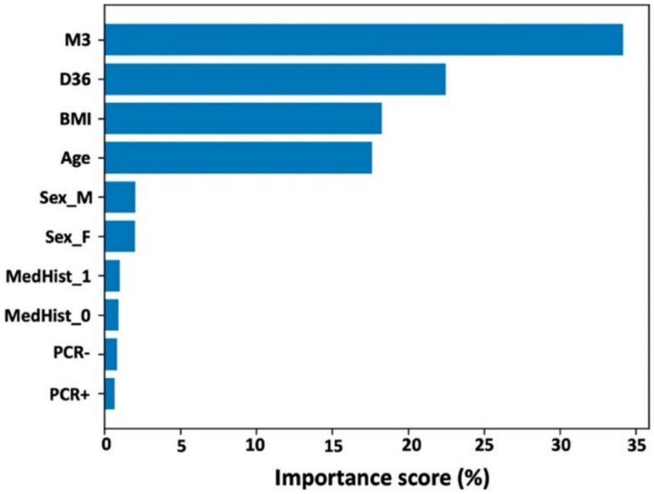


Figure 37. Importance scores for the feature parameters of the subjects participating in the study with regard to their contribution in predicting neutralizing antibody levels nine months after the second vaccination. Key: D36, neutralizing

antibody levels two weeks after second vaccination; M3, neutralizing antibody levels three months after second vaccination; BMI, body mass index; MedHist_1, subjects with autoimmune disorders; MedHist_0, subjects without autoimmune disorders; Sex_M, men; Sex_F, women; PCR+, subjects with previous COVID-19 infection; PCR-, subjects without previous COVID19 infection.

To express how many of the predictions of a random forest classifier were correct and when they were incorrect (i.e., the RF classifier becomes "confused"), a confusion matrix was created (Figure 38).

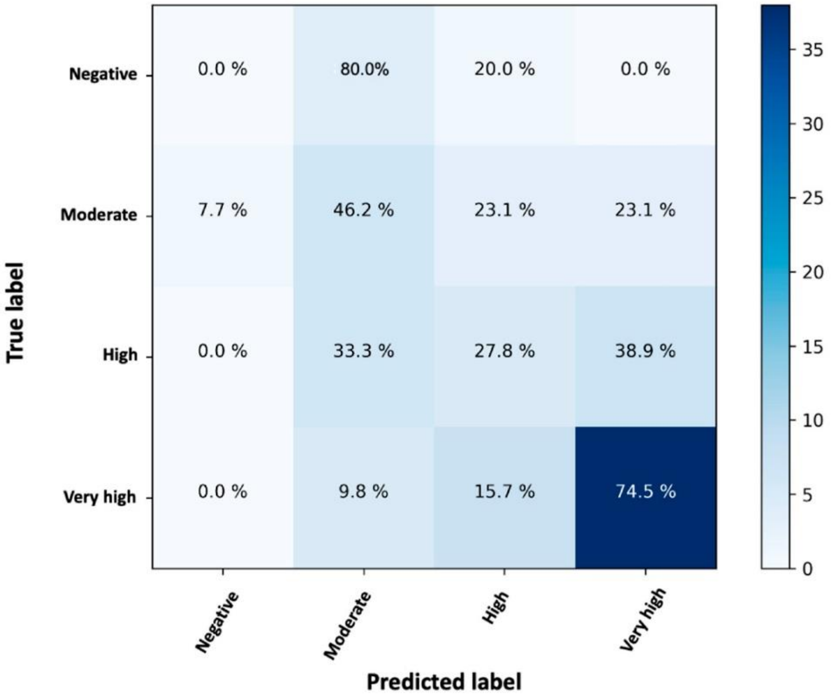


Figure 38. Confusion matrix of the random forest classifier to indicate the percentage of correct or incorrect predictions. The labels "negative", "moderate", "high", and "very high" refer to neutralizing antibody levels (0–30%), (30–50%), (50–75%), (>75%).

The rows of the confusion matrix indicate the observed (true) labels, while the columns represent the predicted (expected) labels. The values within the matrix refer to the percentages of cases assigned to each case, as this is a normalized confusion matrix. Thus, the diagonal values indicate the percentage of cases in which the predicted label matches the observed one. The values in the other cells represent cases where the classifier misidentified an observation. The developed model RF is able to correctly classify 74.5% of very high NAbs values, 46.5% of medium values, and 27.8% of high values (Figure 38). No predictions were made for negative values because there were no subjects belonging to this category.

The prediction of very high values is accurate, but it was less accurate for "medium" and "high" values. The overall prediction accuracy of the RF classifier was 66.32%. This value could be higher if more participants were available for the "negative" group. However, since no data were available for the negative class and most of the data involved very high inhibitions, perhaps the classes were unbalanced and there was a tendency to overpredict. Either way, a percentage of 66.7% (=27.8% + 38.9%) high NAbs were predicted to be either high or very high. Similarly, a percentage of 69.3% (=46.2% + 23.1%) truly moderate cases were predicted to be moderate or high.

C1.4. Discussion

The primal goal of this study was to examine the kinetics of NAbs and anti-S-RBDs against SARS-CoV-2 in 309 healthy individuals, after full vaccination with the BNT162b2 mRNA vaccine for up to 9 months.

A population kinetic model was developed to describe the elimination kinetics of NAbs. The best model consists of a compartment (the whole body), linear kinetics, and a piecewise function for the elimination constant. It is worth noting that 3 different kinetic phases were identified depending on the time period after vaccination. It appears that after the second vaccination, there is a relatively slow elimination of antibodies up to 3 months. It should be mentioned that this apparent slow elimination during the first trimester after vaccination may be due to the fact that NAbs levels are quite high (>90%) and in some cases may exceed the linear phase of the assay. Therefore, this part of the analysis may not be absolutely quantitative. From that point on, elimination increases >6-fold, which means that the NAbs disappear from the body very quickly. Hopefully, this rapid elimination slows down after 6 months and during the last trimester of the study (i.e., from month 6 to month 9), the rate of elimination decreased by about 30% compared with that observed in the second trimester.

To predict NAbs levels at 9, 12, 15, and 18 months after the second vaccine shot, simulations were conducted using the final population model, assuming that NAbs elimination would be the same as found in the last trimester (i.e., from the sixth to ninth month) after the second vaccine. In the future, similar studies involving time points after 9 months (at least one more) would provide a more accurate assessment of the elimination ability of NAbs. In this study, the simulations revealed that the average value of percent inhibition of NAbs would be 59% 1 year after the second vaccination. This value would further decrease to 52% after 15 months and reach a value of 45% 18 months after the second vaccination. These results mean that at 1 year after completion of vaccination, 41% of individuals would not be highly protected. This percentage would be even higher and reach 55% at 18 months.

The significance of our findings stems from the prognostic value of NAbs levels in terms of immunological protection against symptomatic COVID-19 [95]. As a result, the described NAbs reduction in time may inform public health policy for the use of booster doses. Despite the fact that emerging SARS-CoV-2 variants pose a threat to the high rate of immune protection following vaccination, [99, 102] a booster vaccine dose combined with transmission-reducing behaviors remains effective in preventing COVID-19 [103, 104]. It should be noted that in our simulations, we used the identical kinetic parameter values for the extrapolation period after 9 months that we discovered in the 6- to 9-month interval. This option was chosen because it is considered the safest for forecasts, as there are no robust observations on kinetics after 9 months in either the data or the literature. As a result, this option would avoid significant over- or underestimation of NAbs values.

One limitation of this study is the relatively small sample size which can hamper the investigation of specific pathophysiological disorders, such as autoimmune illnesses. Therefore, subgroup analyses should be considered rather exploratory. It should also be mentioned that simulations use the information provided by the experimental data to make predictions. The more representative the original data is of the entire population, the more reliable the simulations will be. A larger sample with more men or patients with several comorbidities would allow for a more accurate investigation and thus more robust simulations. Furthermore, we did not evaluate the kinetics of T-cell responses over time. The sustainability of cellular response against SARS-CoV-2 following vaccination may impact both humoral response and protection against COVID-19 [101, 105-107]. Future research could investigate the neutralizing efficacy of anti-SARS-CoV-2 antibodies against variants of concern as well as the role of anti-S-RBDs toward COVID-19 infection.

The second aim of this study was to use the analytical capabilities of machine learning to determine the parameters responsible for NAbs levels nine months after BNT162b2 vaccination. Data from a clinical trial of 302 subjects were analyzed to uncover correlations between subjects' characteristics that would be difficult or impossible to detect using conventional statistical methods. Figure 39 shows the general route of analysis used in this study.

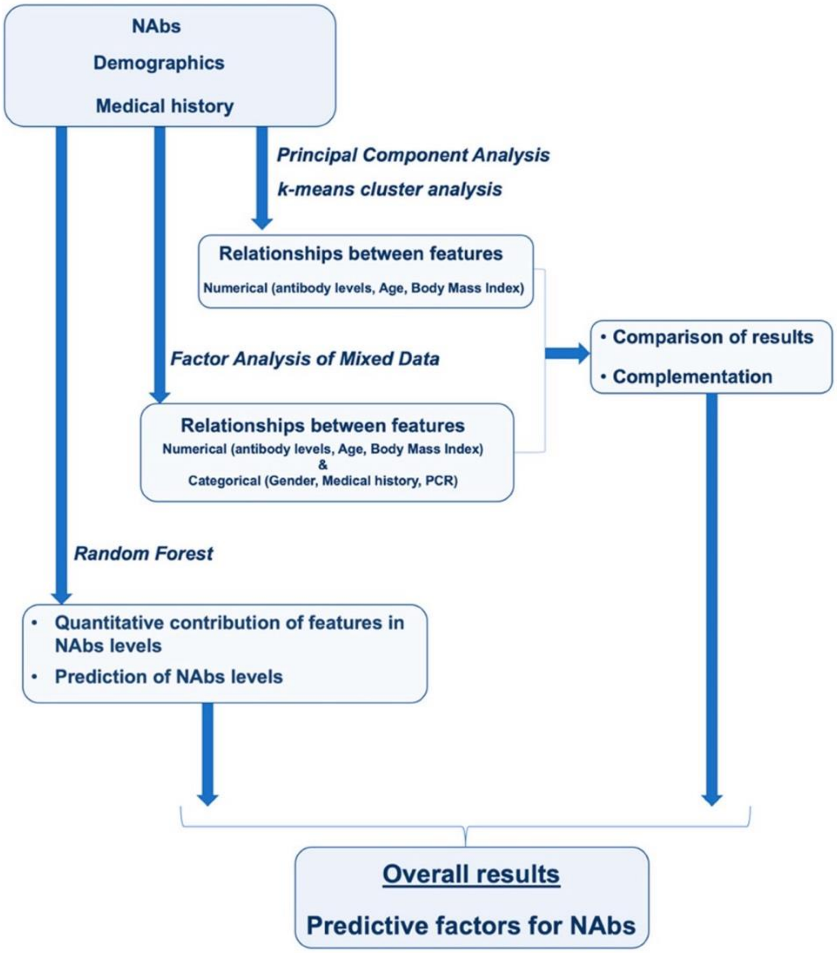

NAbs
Demographics
Medical history
Principal Component Analysis
k-means cluster analysis
Relationships between features
Numerical (antibody levels, Age, Body Mass Index)
Factor Analysis of Mixed Data
Relationships between features
Numerical (antibody levels, Age, Body Mass Index)
&
Categorical (Gender, Medical history, PCR)
• Comparison of results
• Complementation
Random Forest
• Quantitative contribution of features in NAbs levels
• Prediction of NAbs levels
Overall results
Predictive factors for NAbs

Figure 39. Flowchart presenting the overall route of analysis. The association between neutralizing antibody levels (NAbs) and demographics/medical history was investigated using four machine learning methods: principal component analysis, k-means cluster analysis, factor analysis of mixed data, and random forest.

The highest levels of NAbs (median = 97.24%) were observed two weeks after the second vaccination, whereas thereafter there was a statistically significant decline up to nine months. Although the ability to neutralize SARS-CoV-2 antibodies decreased over time, it remained sufficiently high above positive thresholds; a median percent inhibition of 65.43% was observed after nine months. Other prospective studies using classical statistical methods have shown that BNT162b2 provides decreasing but significant COVID-19 protection 6 and 9 months after full vaccination [97, 101,112,113,114]. Data from the field show that protective immunity against the delta version of SARS-CoV-2 decreases after several months of full immunization with BNT162b2 [99,114]. However, protection against COVID-19, particularly against severe disease, remains unchanged compared with unvaccinated individuals [99,114].

Application of PCA (Figure 35) showed the strong relationship between M9 levels and M3 levels and even earlier time points at D36, where maximum inhibition is observed. K-means clustering showed the natural grouping of subjects into five categories. Within each group, the characteristics of some subjects predominate in the sense that one characteristic is distinct from the other clusters. Young subjects, obese subjects, subjects with typical characteristics (i.e., not elderly or obese, etc.), and participants with low NAbs at D36 or M3/M9 make up the five natural groups (Figure 35). The topological characteristics of the groups as well as the PCA vectors can be used to establish their relationships. For example, obese patients have low NAbs, and accumulating data suggest that obesity may attenuate the antibody response to COVID-19 vaccination [115]. Young subjects or those with normal characteristics have higher inhibitory levels. Participants with "typical" characteristics are in the middle, between fat and young subjects. Participants in the last group, with a BMI of 24.83 and an age of 55.99 years, have the best immune response nine months after vaccination.

Factor analysis of the mixed data (Figure 36) showed that younger people have greater amounts of neutralizing antibodies than older people. Based on the observed angles between vectors, age was found to influence NAbs levels more at D36 than at M9. In other words, the influence of age is strongest near the vaccination date, and the role of age appears to decrease with time. This finding is consistent with previous studies (using typical statistical methods) indicating that older age is associated with lower humoral response after SARS-CoV-2 infection or immunization [116-119]. Compared with younger people, the elderly may have a lower capacity to produce antibodies [120]. According to one study, more than one-third of elderly vaccinated subjects lose NAb activity against the delta variant of the pathogen six months after complete immunization with BNT162b2, compared with less than 1% of younger subjects [120]. Although participants' age has a statistically significant effect on NAbs production, as shown by NAbs levels two weeks after vaccination, this effect fades with time.

In addition, FAMD allowed examination of categorical characteristics such as gender, comorbidities, etc. Gender did not appear to affect NAbs levels at M9 (primarily), but there was a modest association with D36. In the latter situation, female subjects (sex = 2 in the graph) appeared to have higher levels of neutralizing antibodies two weeks after the second vaccination. Regarding comorbidities, subjects were divided into two groups: those with autoimmune diseases and those without. FAMD analysis revealed that the vector of the patients with autoimmune diseases was oriented opposite to vectors D36 and M9, meaning that participants with autoimmune diseases had lower inhibitory levels than those without autoimmune diseases. These results are consistent with a previous systematic literature review study that indicated a lower antibody response after COVID-19 vaccination in individuals with rheumatic diseases compared with controls [121]. This feature is more evident in NAbs levels early after vaccination and less evident in NAbs titers nine months after full immunization.

Finally, by using the random forest technique, we were able to quantify the features predicting neutralizing antibody levels nine months after full immunization with BNT162b2. NAbs levels at M3 contribute the most, followed by levels at D36. BMI and age were found to have a small negative impact on M9, while gender had a small impact. Medical history, i.e., whether a person has an autoimmune or other disease (or is healthy), has a small effect on inhibitory levels nine months after full vaccination. Finally, prior SARS-CoV-2 infection before the first vaccination has a small effect on M9 inhibitor levels.

A limitation of this study was the small sample size, which may limit the ability to investigate specific pathophysiologic conditions. Therefore, the subgroup analyzes performed in this study should be considered exploratory. Comorbidities are known to decrease predicted immunogenicity after COVID-19 vaccination; however, this association was small between two weeks and nine months. A larger number of participants in each group is needed to find further significant differences in antibody kinetics. It is also worth noting that patients with serious comorbidities, such as cancer, were excluded from the current study, so the effects of these conditions could not be examined. The role of concomitant medications also could not be studied because the number of subjects in the study was limited compared with the large number of medications that participants were taking. Previous studies have shown that active immunosuppressive drugs can lead to poor humoral response after COVID-19 vaccination [122-126]. However, this aspect could not be investigated in this study. Regarding the role of gender, it should be noted that men and women received unequal sample sizes. The ratio of women to men is approximately two to one. In general, unequal sample sizes can lead to biased comparisons, but we assume that this is not a problem in our case, since the imbalance is not excessive but rather normal (33.8% versus 66.2%). The duration of the study refers to nine months after the second vaccination, almost 10 months after the first dose. Although this observation period is quite appropriate for the current time point, since vaccination programs began almost a year ago, future studies should be conducted with data for a longer period (e.g., 12 months postvaccination) to examine longer-term immune response outcomes.

C1.5. Conclusions

In this study it is shown a sustained but declining humoral immunity against SARS-CoV-2 at 9 months postvaccination with BNT162b2 among 309 healthy individuals. This effect may reflect a declining degree of immune protection against COVID-19 and advocates for the administration of booster vaccine shots especially in areas with emerging outbreaks.

Furthermore, four common machine learning approaches were applied to identify the parameters responsible for NAbs levels nine months after BNT162b2 vaccination. Data from a clinical trial of 302 subjects were analyzed using principal component analysis, k-means clustering, factor analysis of mixed data, and random forest. Application of PCA revealed a strong association of M9 levels with those of M3 and even two weeks after the second vaccination, i.e., when maximal inhibition is observed. K-Means clustering showed the natural grouping of subjects into five categories, with the characteristics of some individuals predominating within each group. These five groupings refer to young subjects, obese subjects, subjects with typical characteristics, and subjects with low NAbs at D36 or M3/M9. Complementary to all these characteristics, factor analysis of the mixed data showed that young subjects had greater amounts of neutralizing antibodies compared with old subjects. Although age has a statistically significant effect on NAbs production, as shown by NAbs levels two weeks after vaccination, this effect decreases with time. In addition, participants with autoimmune diseases were found to have lower inhibitory levels than participants without autoimmune diseases. Finally, we were able to quantify the importance of person characteristics in predicting neutralizing antibody levels nine months after full vaccination using the random forest technique. NAbs levels at M3 contribute the most, followed by levels at D36, whereas BMI and age were shown to have a small negative impact on M9. Gender, previous SARS-CoV-2 infection, and medical history have also been shown to have a small effect on inhibitory levels nine months after full vaccination.

C2. Introducing Data Augmentation in Clinical Studies Using Variational Autoencoders

C2.1 Introduction

Sample size estimation is essential in clinical trials to ensure safety and efficacy [53]. A well-sized sample offers insights into the population but collecting large amounts of data can be challenging and costly. Clinical trials must be meticulously planned with a protocol that includes objectives, endpoints, data collection methods, sample selection criteria, data handling procedures, statistical methods, and a justified sample size [53].

The required sample size depends on the study design, outcome type, and hypothesis test. Key factors in estimating sample size include the minimal detectable difference, variability in measurements, desired statistical power, and significance level [54]. Balancing the sample size is crucial [55]; too small a sample may miss true differences, while too large a sample can be unethical and wasteful. Justifying sample size is increasingly expected by funding agencies, ethics committees, and journals. For highly variable drugs, large sample sizes are often necessary as per regulatory guidelines such as EMA in 2010 and FDA [56,57,58]. Increased variability makes proving bioequivalence more difficult, necessitating larger samples.

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)).

C2.2. Materials and Methods

C2.2.1. 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 no 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 [56,57,58].

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 [56]. 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 [57,58]. 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.

C2.2.2. Variational Autoencoders

Variational autoencoders represent an expansion of conventional AEs [127,128]. 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.

C2.2.3. 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 [129]. 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 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 5 displays all of the aforementioned factors tested during the development of the VAE.

Table 5. 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								

All possible combinations of the factors listed on the left-hand side of Table 5 were investigated. The experimental runs detailed in Table 5 utilized the TensorFlow 2.10.0 Python package and Python version 3.7, with execution taking place within a “Jupyter notebook” environment.

C2.2.4. Monte Carlo Simulations

The methodology used 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 40.

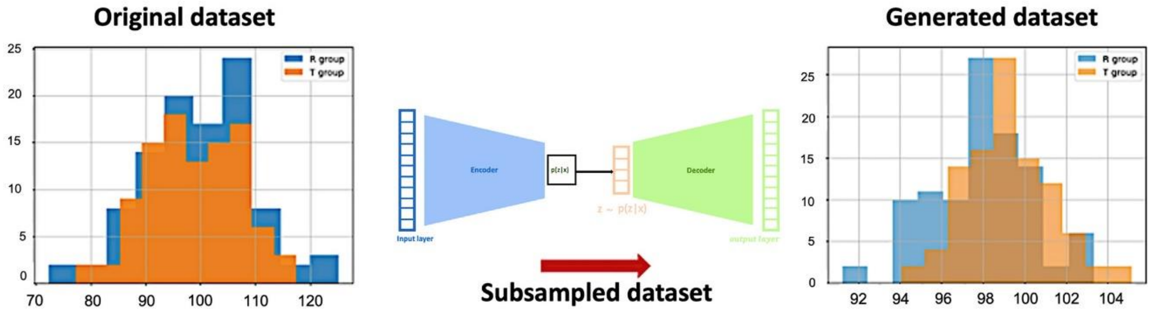


Figure 40. Schematic representation of the analysis strategy in this study. Initially, two randomly generated datasets were generated for the test (T) and reference (R) groups. Then followed subsampling to draw parts of the original population. Finally, the variational autoencoder was applied to the 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.

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×, 1.10×, 1.25×, and 1.50× 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.

After the generation of the virtual subjects, they were appropriately classified into two groups in order to construct a parallel clinical design [56]. 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 [57,58]. Thus, regardless of the distribution of the original data, ln-transformation is always applied in the field of bioequivalence before statistical analysis [56,57,58]. 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% [57,58]. The statistical assessment followed the principles of equivalence testing, namely, the two-one-sided test (TOST) procedure [57,58]. The above-mentioned procedure was repeated several times (500) to allow for reliable estimates.

C2.3. 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 41.

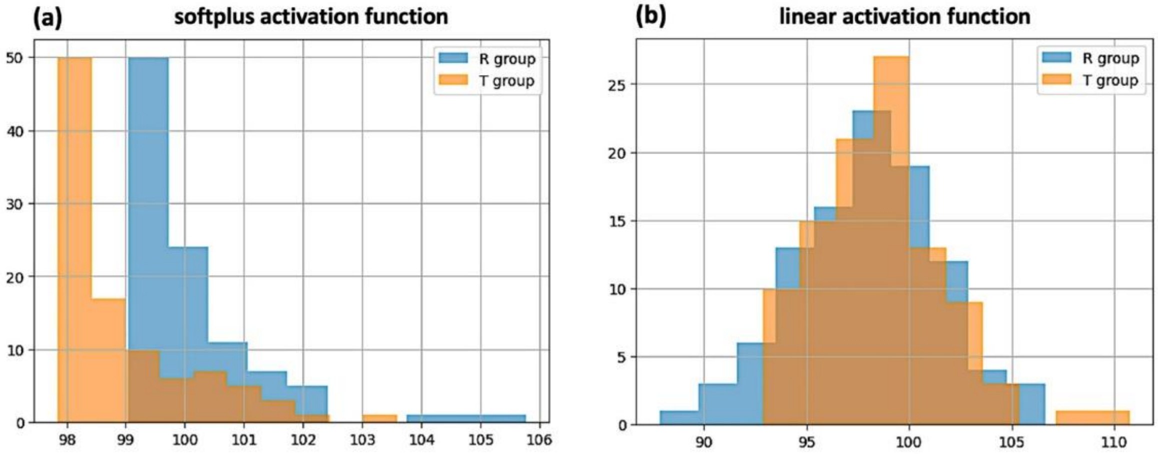


Figure 41. Distribution of the generated data for both R and T groups using the "softplus" (a) and linear (b) activation functions for the output layer.

As illustrated in Figure 41, 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.

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 42.

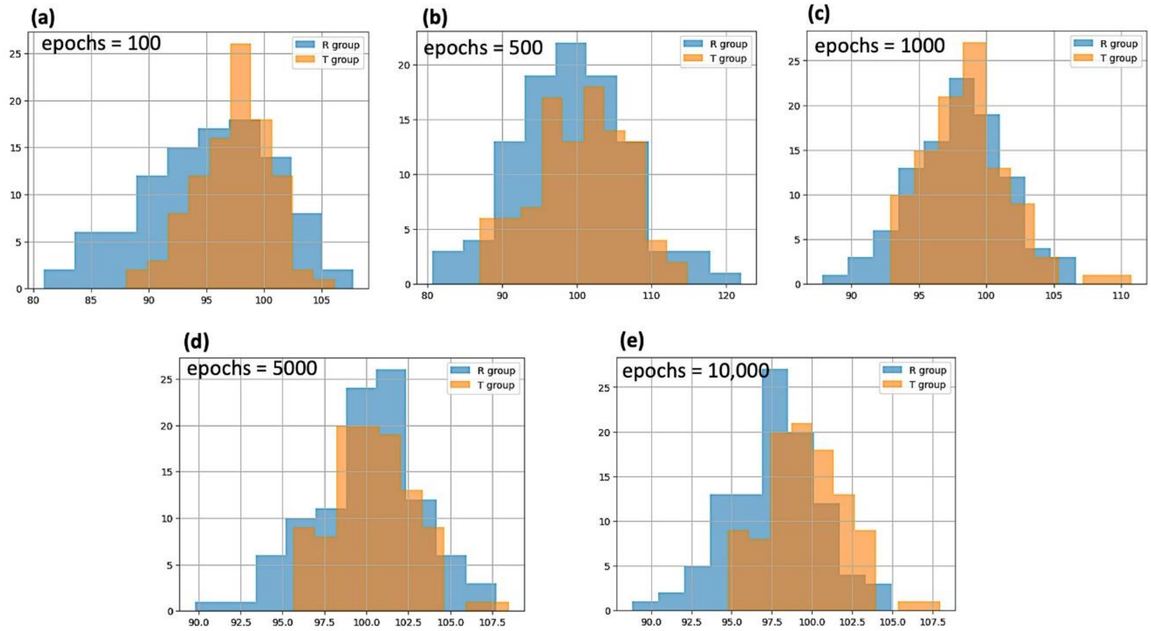


Figure 42. Distribution of the generated data for both the R and T groups using variational autoencoders with 100 (a), 500 (b), 1000 (c), 5000 (d), and 10,000 (e) epochs.

Figure 42 presents a graphical representation that reveals a considerable degree of variance in the data over 100 epochs (Figure 42a). 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 42b). 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 42c). A similar pattern was further observed when the number of epochs was increased to 5000 and 10,000 (Figure 42d,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 43).

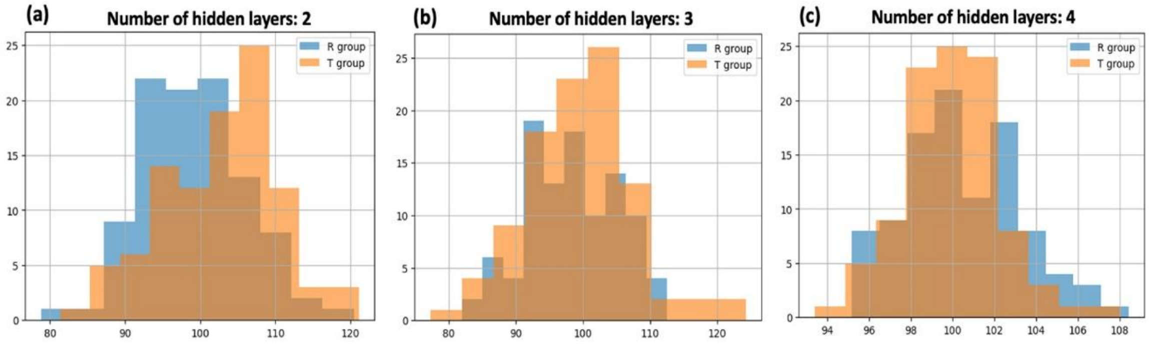


Figure 43. Distribution of the generated data for both the R and T groups using variational autoencoders with 2 (a), 3 (b), and 4 (c) hidden layers for the encoder and the decoder.

Figure 43 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 43a). The findings indicate that there was a noteworthy enhancement in outcomes when the encoder and decoder were equipped with three hidden layers (Figure 43b). 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 43c).

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 44 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.

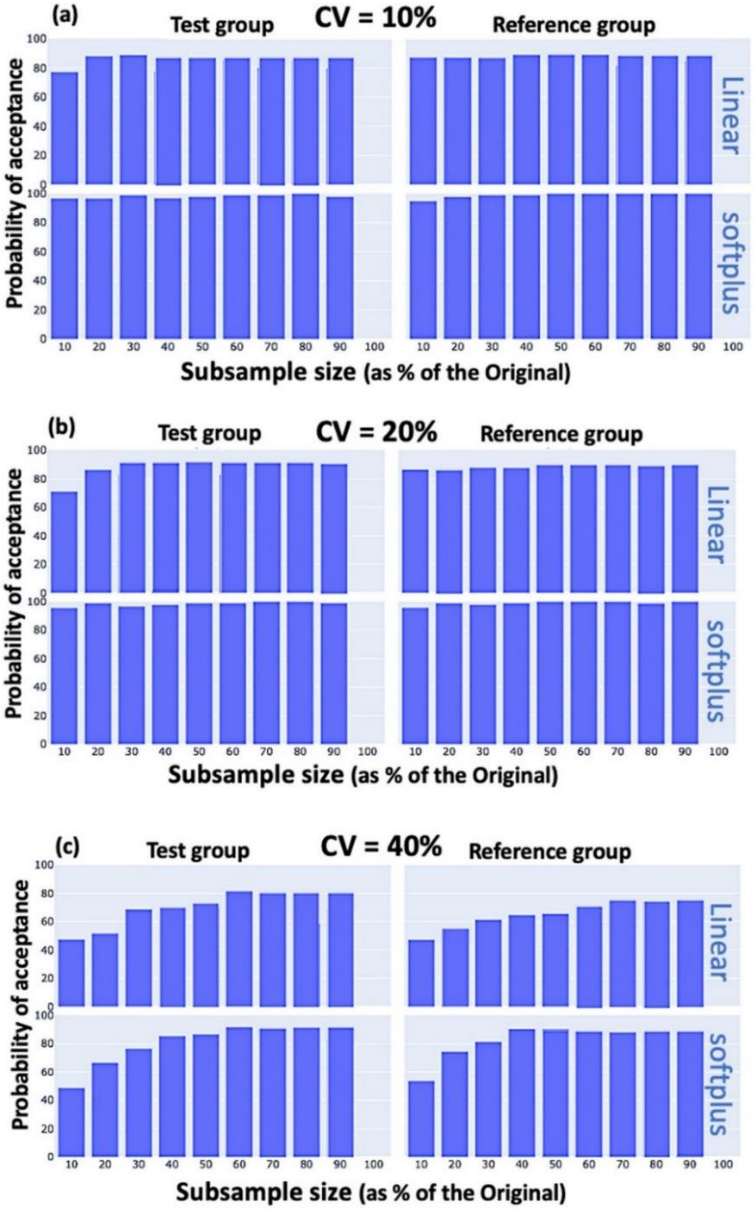


Figure 44. Probability of accepting equivalence between the original and the generated datasets for three levels of variability (CV): (a) 10%, (b) 20%, and (c) 40%. The results are shown separately for the test and reference groups, as well as the two types of activation functions (“softplus” and linear) used for the hidden layers.

Figure 44 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 44a,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 44c). 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.

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 45). 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".

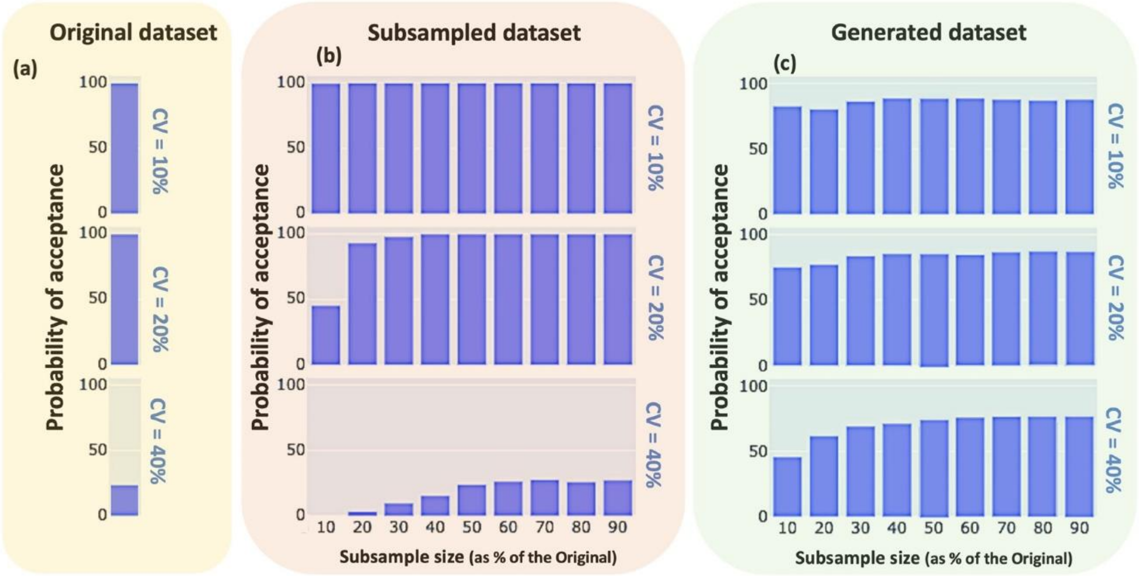


Figure 45. Probability of accepting equivalence between the test and reference groups for the original (a), subsampled (b), and generated (c) datasets Three levels of variability (coefficient of variation, CV) were used: 10%, 20%, and 40%. In all cases, the "softplus" activation was used for the hidden layers, while both the test and reference groups were assumed to exhibit identical average performances.

Figure 45 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 45a), 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 45b) and generated (Figure 45c) 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 45a). 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 45b). On the contrary, the VAE-generated (Figure 45c) 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.

In Figure 45, 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 46 was constructed. In Figure 46, 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, 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.

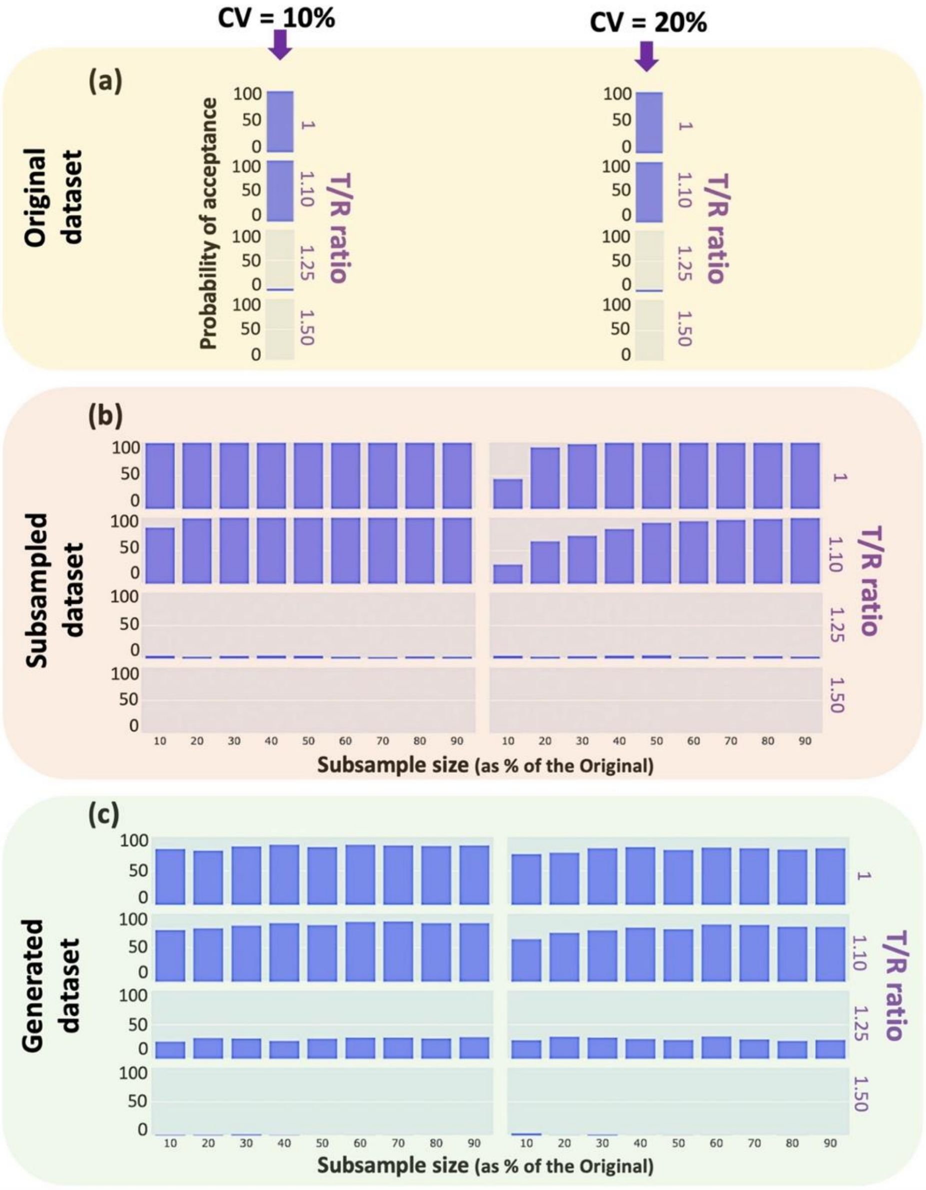


Figure 46. Probability of accepting equivalence between the test and reference groups for several ratios (1, 1.10, 1.25, 1.50) of the average test (T)/reference (R) performance. The comparisons were made separately for the original (a), subsampled (b), and generated datasets by the variational autoencoder (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%.

In all cases in Figure 46, a decrease in statistical power can be observed with an increase in the T/R ratio. For the original data (Figure 46a), 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 46b). 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 46c), 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 46c, 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.

C2.4. 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 [129].

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 (Figure 44, Figure 45 and Figure 46). It is noteworthy that for high variabilities (Figure 44c), the performance of the VAE method became even better than with the original dataset.

he 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 [129].

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 41). Subsequently, a determination had to be made regarding the appropriate number of epochs used in the VAE. As illustrated in Figure 42, 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 43).

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 (Figure 45b and Figure 46b). On the contrary, the generated data using the VAE showed increased statistical power when the data exhibited either low or high variability (Figure 45c and Figure 46c). 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 45c). 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 45c). 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 [56]. 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., Cmax), using different statistical methods to evaluate bioequivalence (e.g., scaled equivalence), or increasing the number of samples in the study [56]. 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 46). 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 46c). These results are consistent with the existing literature in the field of highly variable drugs [56]. 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 44).

C2.5. Conclusions

The use of neural networks, particularly variational autoencoders, was introduced, as a tool to virtually increase the sample size in clinical studies and thereby decrease the required number of actual participants. Firstly, the most appropriate architecture for the VAE and tuning hyperparameters was developed, 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. 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.

C3. On the Use of Variational Autoencoders for Generating Virtual Subjects in Bioequivalence Studies

C3.1. Introduction

Estimating sample size in clinical trials is a crucial step that demands careful attention, as errors in this calculation can lead to misleading conclusions and jeopardize the trial [53,130]. Sample size determination is affected by factors such as the study design, type of outcome, statistical hypotheses, anticipated measurement variability, minimum detectable difference, statistical power, and significance level [54]. Similar considerations apply to bioequivalence studies, particularly when comparing a generic pharmaceutical product (Test, T) to a reference product (R). Two pharmaceutical products are considered bioequivalent if they contain the same active substance at the same molar dose and demonstrate equivalence through comparative pharmacokinetic studies, known as bioequivalence trials. Once bioequivalence is confirmed in these trials, the products can be regarded as therapeutically equivalent [57,58].

In this context, the concept of using generative AI algorithms in clinical trial data augmentation is introduced. This work investigates the suitability of using VAEs to virtually increase the sample size in the typical situation of BE studies, namely, the two-period, two-treatment 2 × 2 crossover design with a washout period [131].

A computational methodology that combines Monte Carlo simulations of 2 × 2 BE trials with deep learning methods (i.e., VAEs) is developed. Various scenarios, including variability levels, the actual sample size of the BE study, AI-generated sample size, and the relationship in the average performance between the T and R pharmaceutical products, were explored. Our ultimate purpose was to assess the usefulness of VAE as an AI method in bioequivalence studies, aiming to significantly reduce human exposure, costs, and trial completion time.

C3.2. Materials and methods

To assess the utility of VAE in BE studies, it was crucial to replicate the actual conditions of BE testing. The main components of the methodology include generating virtual subjects using Monte Carlo simulations, training/tuning the VAE model to create “synthesized” subjects, applying typical BE testing conditions imposed by regulatory authorities (such as the appropriate statistical framework and acceptance limits), and repeating the entire process several (i.e., hundreds) times to obtain robust estimates. The aforementioned procedure was applied across various scenarios (e.g., different T/R ratios, variability, and original sample size).

C3.2.1. Variational autoencoders and tuning of hyperparameters

To accomplish this task, variational autoencoders were leveraged and tuned. We extensively fine-tuned the hyperparameters, which included searching for the most suitable activation functions (e.g., softplus, linear, ReLU, sigmoid, etc.), determining the optimal number of hidden layers, and adjusting the number of epochs. The softplus activation function performed best for hidden layers, while the linear function was optimal for output layers. Also, the optimal number of hidden layers and epochs was 3 and 1000, respectively. Similarly, the optimal quantity of neurons for every hidden layer, from left to right, was found to be 64-32-16 for the encoder and 16-32-64 for the decoder. Additionally the impact of unequally weighting the two parts of the cost function was explored and standardizing the input data by removing the mean and scaling to unit variance. Table 6 summarizes all the combinations of settings that were investigated:

Table 6. Hyperparameter values explored in the tuning process of the VAEs. Across all instances, the latent space dimension and the number of epochs were fixed at 1 and 1000, respectively. Standardizing the input data was applied to all cases.

Activation Function		Weights of Loss Function		Number of Hidden Layers		Number of Neurons in Hidden Layers	
Hidden Layers	Output Layer	Kullback–Leibler Part	Reconstruction Part	Encoder	Decoder	Encoder	Decoder
Softplus	Linear	1	1	3	3	64-32-16	16-32-64
		2	2				
		…	…				
		9	9				
		10	10				

C3.2.2. Simulation framework

The simulation of the bioequivalence in the context of 2 × 2 crossover design consisted of generating subjects for the reference (i.e., R) and the test (i.e., T) groups for both periods, termed as “original” [56]. Initially, a total of N_R subjects was generated for the first period, corresponding to the R product, from a random process with mean $\mu_{R1}$and standard deviation σ_{R1}. For the T

drug, N_T subjects, equal to N_R, were generated, following the same process with mean $\mu_{T1}$and standard deviation σ_{T1}. Both groups had equal coefficient of variation (CV), referring to the between-subject variability. For the second period, each of the total $N = N_R + N_T$ subjects was multiplied with a stochastic term, which was generated through a random process, to incorporate the within-subject variability of the subjects.

The original data were randomly subsampled with varying proportions, referred to as "subsampled". These subsampled datasets were used to train the VAE model and, in a next step, to synthetize the new virtual subjects termed as "generated". Finally, the "generated" data from both groups (T and R) and study periods (i.e., periods I and II of the 2 × 2 crossover design) were transferred for statistical analysis, using the official statistical framework imposed by regulatory authorities for bioequivalence [57,58]. This statistical analysis refers to the following steps: ln-transformation of the variables, application of a linear model (ANOVA), calculation of the residual error, using this error term to construct a 90% confidence interval for the mean difference between T and R in the ln-domain, and, lastly, checking if BE is declared [56,57,58]. This entire procedure was repeated 500 times and the percentage of bioequivalence acceptance was measured at the end [56].

The above-mentioned route of analysis can be outlined in the following steps:

I. N individuals are randomly generated for both groups in the case of the first period:
    a. N_T individuals for the T group for the first period, with mean $\mu_{T1}$and standard deviation σ_{T1}.
    b. N_R individuals for the R group for the first period, with mean $\mu_{R1}$and standard deviation σ_{R1}.
    c. The T and R groups were set to have equal CVs;
    d. The sample sizes of T and R groups were assumed to be equal: $N_T = N_R$.
    e. Thus, the sample size of the study is: $N = N_T + N_R$.

II. The N individuals from the first period are multiplied, with a randomly generated "stochastic term", with mean μ_{ST} and standard deviation σ_{ST}. The stochastic term coefficient of variation (CVw) represents the "within subject variability" for each simulated volunteer between the period I and II of the crossover study [56]. It should be noted that the CVw is different to the between-subject variability (i.e., CV) discussed in step "Ic".

III. The individuals generated from steps I and II (termed as “original”) are then randomly subsampled with proportions 25%, 50%, and 75%. The so-derived groups are termed as “subsampled”.

IV. Subsampled individuals are fed into an optimized VAE model to generate new individuals, termed as “generated”. The generated dataset was set to exhibit size equal to or double the “original” dataset.

V. The standard statistical criteria mandated by regulatory authorities are utilized to assess BE among all comparison groups [57,58].

VI. The success (i.e., BE acceptance) or failure (i.e., non-equivalence) of the statistical test is tracked for all three datasets.

VII. Steps “I–VI” are performed again for 500 repetitions in order to obtain robust values for the % BE acceptance.

VIII. The results attained from step “VII” are evaluated.

A list of factors analyzed in this study, including CVw, the ratio between $\mu_{T1}$and μ_{R1}, N, subsampled proportions, and the size of generated data proportionate to the total size N, is presented in Table 7:

Table 7. List of factors explored in this study. For all cases, the mean of the stochastic term, mentioned in step “II” above, was set equal to 1.

Between-Subject Variability (CV)	Within-Subject Variability (CVw)	Mean Endpoint Value for the Reference	Ratio of Average Endpoints Test/Reference	Original Sample Size (N)	Subsampled Proportions	Size of Generated Data (xN)
20%	15%	100	1	12	25%	1×
	30%		1.1	24	50%	2×
			1.2	48	75%	
				72		

C3.3. Results

The first exploration in this study, referred to the condition when both the T and R pharmaceutical products have equal means and a low coefficient of variation at 15%. This scenario was investigated for various sample and subsample sizes (Figure 47).

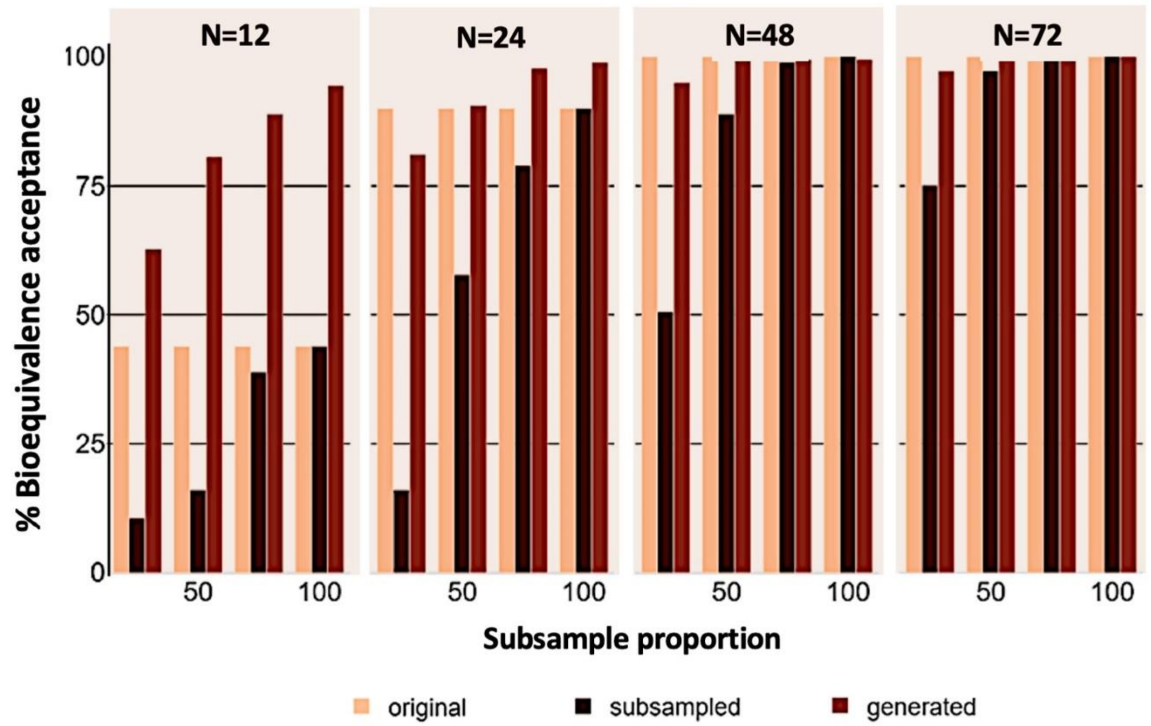


Figure 47. Bioequivalence acceptance rate (%) for the "original", "subsampled", and VAE-"generated" datasets when both the Reference (R) and Test (T) pharmaceutical products exhibit identical average performance. Four different "original" sample sizes (N) were utilized (12, 24, 48, and 72). The subsample proportions were 25%, 50%, 75%, and 100%.

As illustrated in Figure 47, the BE acceptance rate increases with the rise in the actual sample size. In addition, the percentage of acceptance for the "subsampled" group is always lower than the "original", whereas the acceptance rate of the "VAE-generated" group of subjects is always higher than the original and, thus, the "subsampled" also. Also, as expected, for all the three datasets (original, subsampled, and VAE-generated), the acceptance rate tends to decrease when the subsample proportions decline or when the sample size gets lower. Though, for the original and the subsampled datasets, this decrease is more extreme than the generated. More specifically, for the "generated" dataset, the acceptance rate starts from 60% when the sample size and the subsample proportion are small and increases up to 100% while the sample size and subsample proportion increase. This trend is similar for the original and subsampled datasets, though the acceptance rate for the original dataset reaches its lowest (around 40%) for sample size 12, while, for the subsampled dataset, the lowest is reached even at the sample size of 24 for low subsample proportions (around 10%).

Figure 48 illustrates the probability of acceptance for the same sample sizes and proportion rates between the R and T groups for low CVw (15%) when the mean of the T group is 10% and 20% higher than the one of the R group (i.e., for T/R ratios of 1.1 and 1.2).

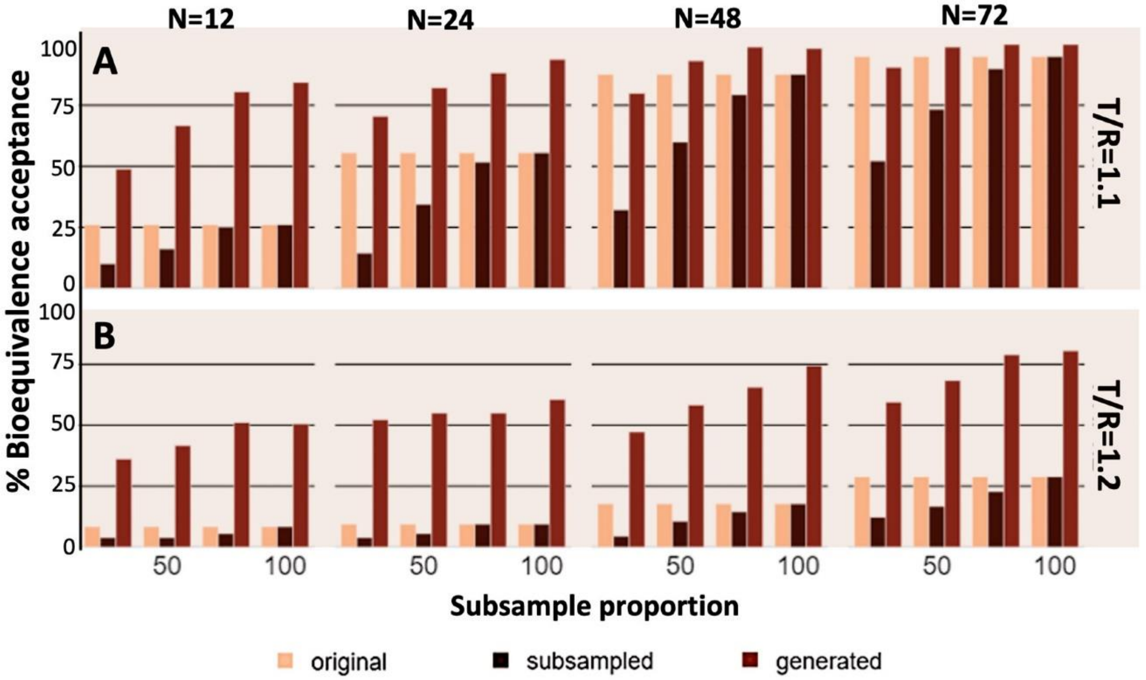


Figure 48. Bioequivalence acceptance rate (%) for the "original", "subsampled", and VAE-"generated" datasets when there is 10% (A) and 20% (B) difference in the average endpoint value between the Test (T) and Reference (R) pharmaceutical products. Four different original sample sizes (N) were utilized (12, 24, 48, and 72). The subsample proportions were 25%, 50%, 75%, and 100%.

Figure 48 illustrates the trend of BE acceptance as both sample size and subsample proportions increase from left to right. This analysis is conducted with the T/R ratio of the average endpoint set at 1.1 in the top row and 1.2 in the bottom row. Notably, the acceptance rate of the generated data is significantly higher than that of the original and subsample data. This difference is particularly evident when both the sample size and proportion rate are lower, as shown in Figure 3A,B. As expected, a larger difference between the means of R and T groups corresponds to a lower acceptance rate across all groups. It is essential to emphasize that the acceptance rate of the generated data consistently exceeds that of the other two datasets.

Figure 48 similarly illustrates the acceptance rates for a 30% within-subject variability of the measured BE endpoint and several values of the T/R ratio (1.0, 1.1, and 1.2).

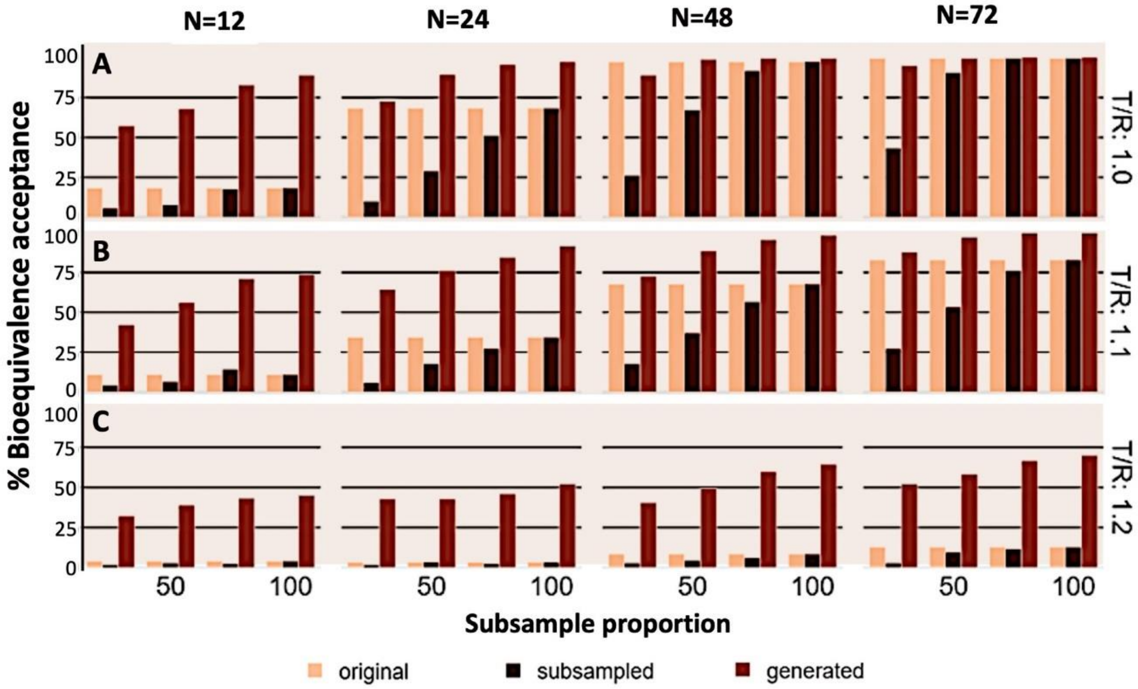


Figure 49. Bioequivalence acceptance rate (%) for the "original", "subsampled", and VAE-"generated" datasets for high variability values (CVw = 30%). The T/R of the Test (T) and Reference (R) products was set at 1.0 (A), 1.1 (B), and 1.2 (C), while the subsample proportions were 25%, 50%, 75%, and 100%. Four different original sample sizes (N) were utilized (12, 24, 48, and 72).

In Figure 49, a consistent trend is evident across all scenarios examined. Again, it should be emphasized that the BE acceptance rate of the generated dataset is consistently higher than that of both the subsampled and original datasets in all cases. Additionally, as expected, the BE acceptance rate decreases for all groups with the increase (or decrease) in the T/R values. Importantly, even in this scenario, the percentage of BE acceptance for the generated data is the least affected. Figure 51 shows the impact of within-subject variability of CVw on BE acceptance "gain" between the generated and original datasets (i.e., % BE acceptance generated—original). The scenarios explored referred to different performances between the two drug products under comparison, namely, T/R values of 1.0 (Figure 50A), 1.1 (Figure 50B), and 1.2 (Figure 50C).

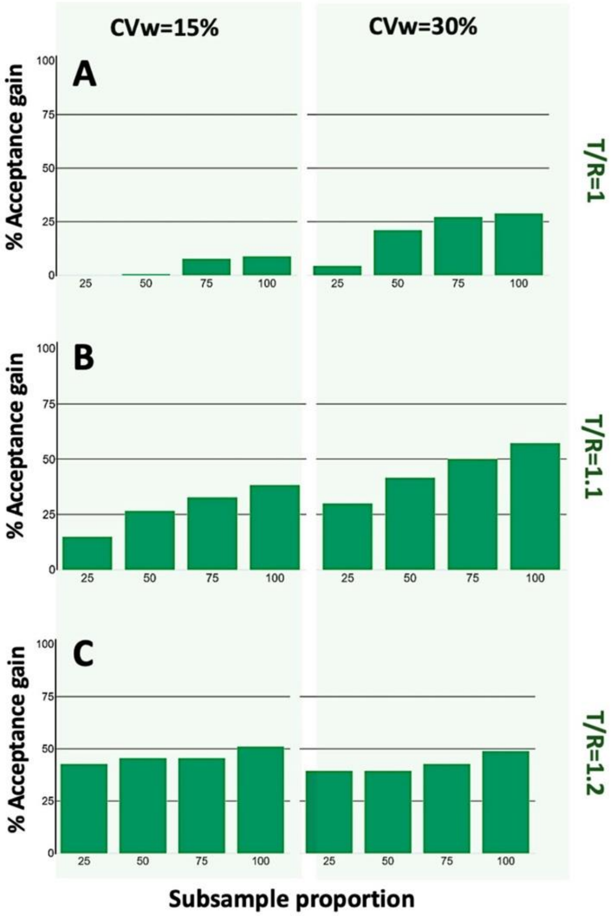


Figure 50. Bioequivalence acceptance "gain" of the VAE-"generated" dataset minus the "original" datasets. Three Test/Reference (T/R) ratios of the average endpoint were considered: 1.0 (A), 1.1 (B), and 1.2 (C). In all cases, a range of subsample proportions (25%, 50%, 75%, and 100%) was investigated, while the within-subject variability (CVw) was set at either 15% (typical case) or 30% (highly variable situation).

Figure 50 illustrates that, in all cases, the acceptance rate of bioequivalence between the T and R groups is significantly higher in the case of the generated dataset. When the T and R groups have equal means (Figure 50A), this increase is more pronounced for larger within-subject variability and larger subsample proportions. Similar findings are observed when the T/R ratio is 1.1, with the increase being even higher (Figure 50B). For more extreme differences between the T and R groups (T/R = 1.2), an even more substantial increase in the acceptance rate is observed, although the increase is very similar for both 15% and 30% CVw and for all subsample proportions (Figure 50C).

Finally, Figure 51 investigates the relationship between the generated sample size and the original dataset. In other words, the aim of this figure is to show how many times larger the generated dataset can be compared to the original one. Two typical N values were chosen: N = 12, which refers to the lowest accepted sample size for a BE study by regulatory authorities [57,58], and N = 24, which is a typical sample size commonly used in BE studies. Two levels of the generated sample size were tested: 1× and 2×. This means that the generated dataset can be either as large as the original or twice the original sample size. Thus, by using only part of the original dataset (i.e., 25%, 50%, or 75%), the aim was to evaluate the performance of the reconstructed dataset. In all cases, two CVw values were utilized (15% for Figure 51A and 30% for Figure 51B) and three levels of the relationship between the T and R groups (i.e., T/R equal to 1.0, 1.1, and 1.2).

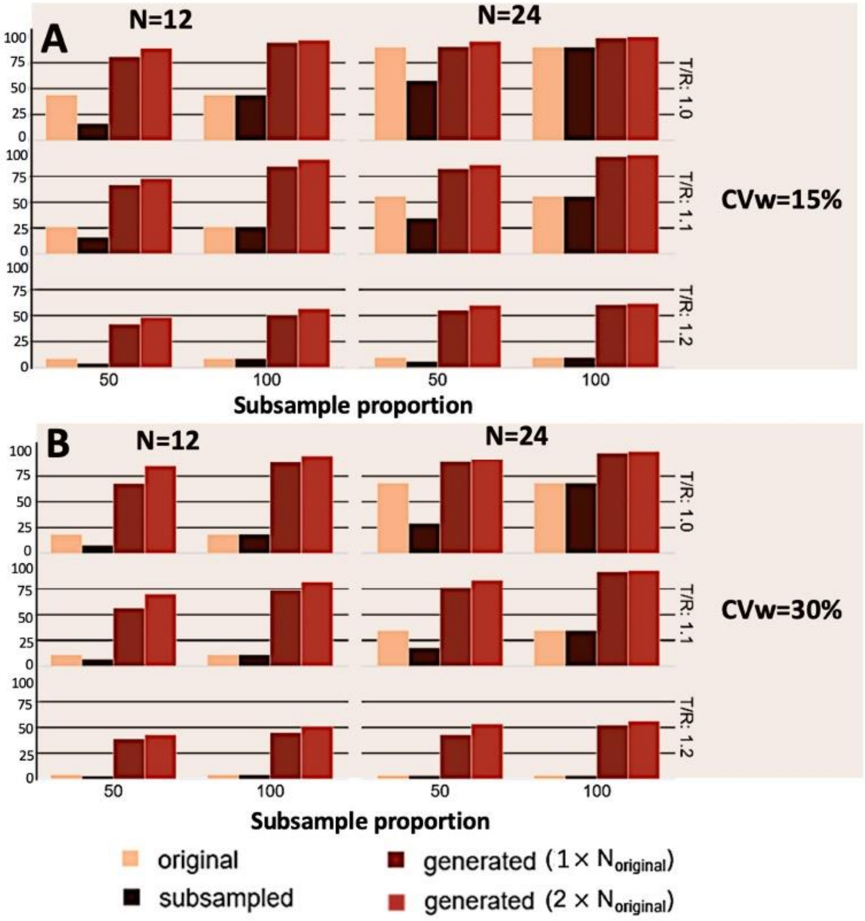


Figure 51. Bioequivalence acceptance of the VAE-"generated", "original", and "subsampled" dataset for within-subject variability of 15% (A) and 30% (B). Two possibilities of the VAE-"generated" datasets are shown: one with the same sample size as the "original" dataset and another with twice the "original" sample size. Three Test/Reference (T/R) ratios of the average endpoint were considered: 1.0, 1.1, and 1.2. In all cases, a range of subsample proportions (25%, 50%, 75%, and 100%) was investigated.

Figure 51 demonstrates that the acceptance rate increases when the sample size of the generated dataset is twice the original size, compared to cases where the original sample size is equal to the generated sample size. In Figure 51A, when CVw is 15% and the original sample size is 24, a slight increase in the acceptance rate is observed for a subsample proportion of 50% across all T/R ratios. When the original sample size is 12, there is a modest increase when the generated sample size is twice the original for a subsample proportion of 100% and a more significant increase when the subsample proportion is 50%. This pattern is evident for all three T/R ratio values. In Figure 51B, where CVw is 30%, the results mirror those of Figure 51A. However, since CVw is higher, the percentage of BE acceptance is lower in all scenarios. The same findings are also observed in the case of a highly variable drug (Figure 51B). In this instance, the superior performance of the two VAE-generated datasets is even more evident.

C3.4. Discussion

In this study, the concept of utilizing VAEs in clinical trials is extended to the field of bioequivalence studies. So, this study aims to explore how the utilization of a VAE-generated dataset can be used to lower the need of actual human data and substitute them with AI-generated data. To accomplish this task, we simulated the conditions of 2 × 2 (i.e., two-period, two-treatment) crossover BE trials. Initially, a group of original data (i.e., study volunteers) was simulated, and synthesized data were created by taking parts of them (e.g., 25%, 50%, 75%, or all of them, i.e., 100%) using the VAE model. The entire procedure was repeated multiple times through Monte Carlo simulations, exploring various conditions (scenarios). These conditions included within-subject variability, different average endpoint values for the Test and Reference products (i.e., the T/R ratios), subsampled proportions, and how many times the generated data were larger than the original (i.e., 1×, 2× the original dataset). These scenarios were selected to explore various conditions of special interest. For instance, a T/R ratio of 1 corresponds to the case where the two groups under comparison exhibit identical performance, resulting in a relatively high percentage of bioequivalence acceptance. In contrast, the case of T/R = 1.1 implies a 10% average difference between the two groups, leading to reduced BE acceptances. Also, the role of within-subject variability (i.e., CVw) is crucial in bioequivalence studies, as the statistical assessment relies on estimating a 90% confidence interval where the variability term used is CVw. Thus, in this study, two levels of CVw were examined: a moderate value of 15% and a marginal value of 30%, which is the threshold for a pharmaceutical product to be considered highly variable. It is worth mentioning that an advantage of using VAE neural networks lies in their inherent ability to reduce the variability of input data. This is particularly important for BE studies, especially in the case of topical pharmaceutical products. In fact, additional scenarios involving highly variable drugs and/or pharmaceutical products should be studied, but this could be the focus of an entirely new study. Additionally, this study investigated the robustness of AI-driven generation of virtual subjects concerning the amount of data used. To address this, subsamples as low as 25% of the original dataset were examined, demonstrating the desired performance. Thus, after tuning the VAE hyperparameters, the performance of the VAE system was evaluated in terms of the percentage acceptance of bioequivalence compared to the original and subsampled datasets.

Given that the BE limits are 0.80–1.25 and the T/R ratios used in this study were narrower (i.e., 1.0, 1.1, and 1.2), the desired outcome would be a situation where the percentage acceptance of the generated dataset is higher than that of the original data and certainly much higher than the subsampled data from which they were created. In all scenarios studied, it was demonstrated that the VAE system can successfully generate data superior to any subsampled group and exhibit performance at least equal to the original dataset (Figure 47, Figure 48, Figure 49, Figure 50 and Figure 51). To clearly emphasize this advantageous performance of the VAE model, Figure 5 was constructed to illustrate that, in all cases, the VAE-generated data lead to at least equal or better performance compared to the original data. As expected, the reduced sample dataset (referred to as “subsample”) exhibited inferior performance compared to the original dataset. The subsampled data failed to reproduce the characteristics of the original dataset, lacking the necessary statistical power to establish bioequivalence, particularly when it was present, as a result of the diminished sample size (Figure 47, Figure 48 and Figure 49). However, what is crucial is that relying on this “inferior” dataset (i.e., the “subsampled”) and applying the VAE model to it leads to synthesized data that exhibit performance at least equal to the original whole data. In certain instances, the efficiency of the data generated by the VAE is notably superior, even compared to the original data.

For example, with low within-variability (15%) and when the mean endpoints of the reference and test groups were the same (Figure 50A), the data generated by the VAE exhibited comparable performance to the original dataset, even when only a small portion of the original data was utilized (i.e., 25%). In other words, only a few actual human subjects were included in the study. Generating even more data than the original will lead to an even higher acceptance rate (Figure 51).

The VAE-generated data performed well, even in cases of high variability and when the mean endpoints between the groups differed by 20%, resulting in an increase in statistical power (Figure

48 and Figure 49). It was demonstrated that an increase of up to 40% in statistical power can be achieved when the mean endpoints of the two groups have a 20% difference, even with a 25% subsample proportion, namely, when only a few actual volunteers will participate in the trial (Figure 50C). It should be underlined that high variability poses a significant challenge in the field of BE assessment [56,57,58]. However, achieving strict BE for such drugs can be daunting due to their inherent pharmacokinetic variability. As a result, regulatory bodies have developed particular guidelines and criteria for the acceptance of highly variable drugs to address the expected variability [56,132]. Therefore, employing methods that mitigate unwanted variability without necessitating an increase in the number of study participants, costs, or study complexity becomes paramount.

One limitation of this work is the relatively low number of repetitions utilized for each scenario. Due to computational constraints, completing 500 runs consumed a significant amount of time; however, the estimates appeared robust and only minimally different compared to 200 or 300 runs, implying that convergence was achieved with 500 runs. Further exploration in this field could include the assessment of more scenarios such as additional clinical designs (e.g., replicate) and statistical hypotheses (e.g., noninferiority or superiority). However, it was impossible to investigate all these possibilities encountered in clinical trials.

C3.5. Conclusions

This study is our second step toward utilizing generative AI algorithms in clinical trials. The typical conditions of 2 × 2 crossover BE studies were simulated by combining Monte Carlo simulations with VAEs. Various scenarios, including variability levels, the actual sample size of the BE study, the AI-generated sample size, and the relationship in the average performance between the T and R products, were explored. All simulations performed in this study showed that incorporating AI generative algorithms in clinical trials has many advantages in creating virtual populations. These advantages include a decrease in human exposure, significantly shorter study completion times, lower complexity in the clinical trial, reduced workload for physicians and clinics, and significantly lower costs for sponsors or health agencies. It was shown that less actual human data can be used to achieve similar, and even better, results in terms of statistical power. Overall, this study suggests the utilization of AI-driven generative algorithms in clinical research. However, the incorporation of such new ideas in practice would require regulatory authorities to set specific criteria and guidelines on the minimum requirements for the application of AI-generated virtual subjects in order to avoid possible pitfalls (e.g., hallucinations) and ensure reproducibility.

C4. Bioequivalence Studies of Highly Variable Drugs with the Use of Variational Autoencoders

C4.1. Introduction

‘Highly variable drugs’ are those with a within-subject coefficient of variation of 30% or more, due to either the drug substance or its formulation [56]. 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.

When applying bioequivalence assessment for the comparison of the pharmacokinetic properties of two drug products, the typical average BE, falls short. This is due to the fact that BE is typically confirmed, if the 90% confident interval falls within the established range of 80-125% [62,63]. Variability can stem from factors such as drug characteristics or patients’ physiological conditions. As within-subject variability increases, demonstrating BE becomes more challenging without increasing the sample size. To address this, various methods have been proposed. Both the EMA and the U.S. FDA recommend using scaled BE limits, which adjust the BE limits based on the within-subject variability of the reference product. These reference-scaled procedures, recommended by the EMA and FDA, require full-replicate or semi-replicate study designs, where the reference product is administered at least twice to each subject, allowing for accurate estimation of within-subject variability [62,63].

C4.2. Materials and Methods

C4.2.1. 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 [56]. 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 52).

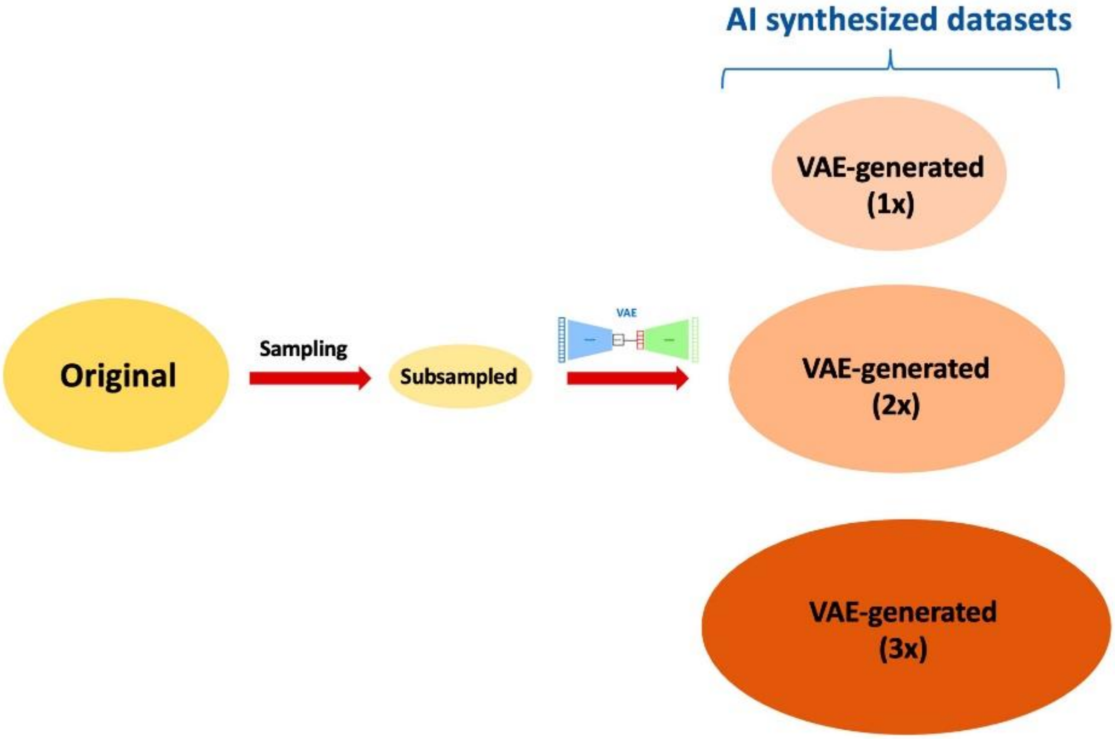


Figure 52. A graphical illustration of the general idea for using VAEs as a tool for data augmentation in bioequivalence studies with high-variability drugs. Instead of requiring a large sample size (termed “original”), only a subgroup of this (i.e., “subsampled”) is needed. VAE is then applied to the 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 [62,63]. 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

[56]. These studies have demonstrated that conducting this quantity of repetitions ensures the acquisition of reliable and consistent estimates [56].

C4.2.2. 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 [133]. 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 8. 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 8. 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 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.

C4.2.3. 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 [56,62,63]. 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 [62,63].

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 μR1 and standard deviation σR1 for the R group and mean μT1 and standard deviation σT1 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 σw. This stochastic term (with coefficient of variation CV_W) represented the within-subject variability [56]. 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 generated 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 9:

Table 9. 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				

C4.3. Results

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.

Figure 53 illustrates the BE acceptance rates for different coefficients of variation of 20% (Figure 53A), 40% (Figure 53B), and 60% (Figure 53C). 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 [56].

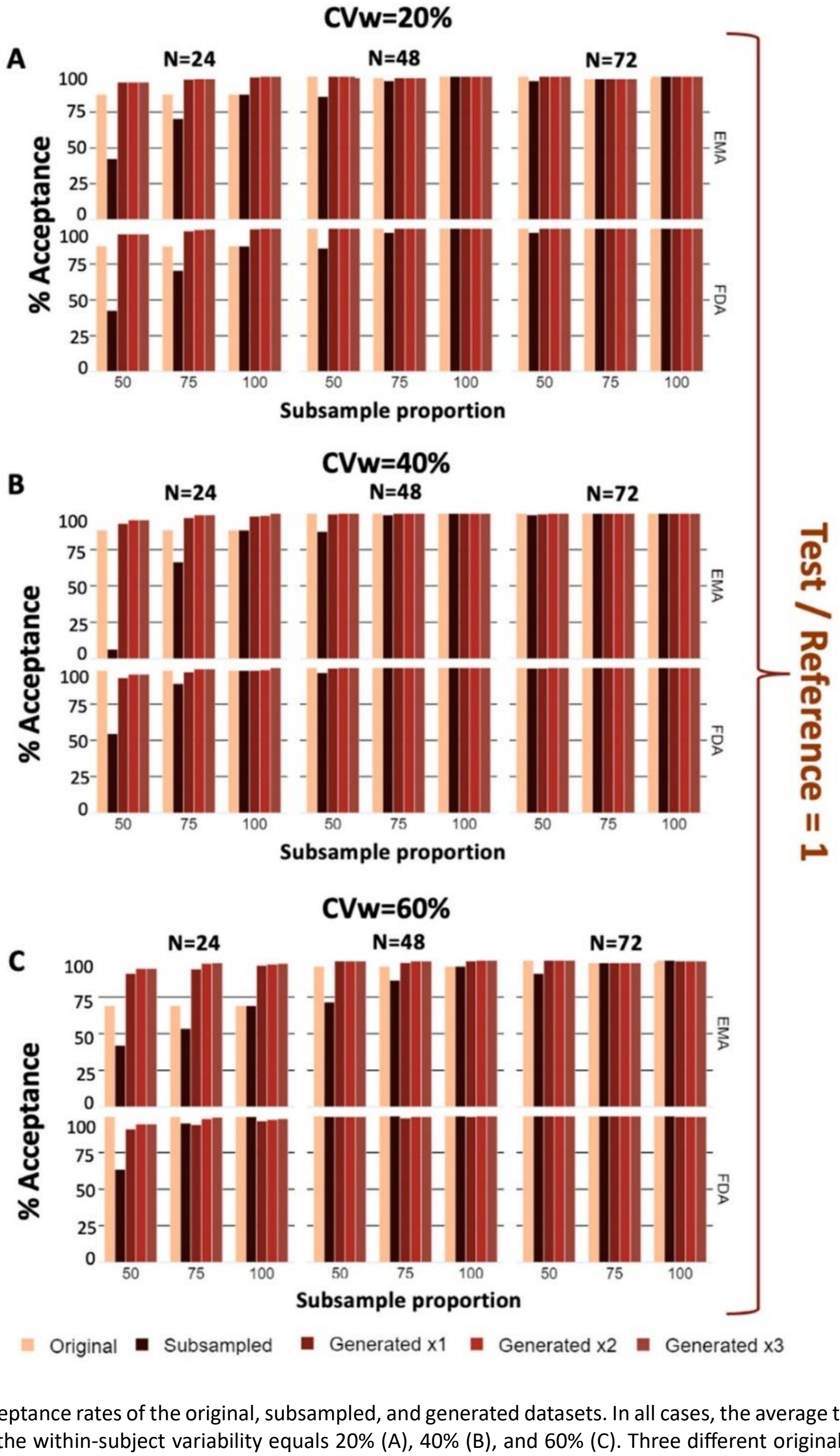


Figure 53. Acceptance rates of the original, subsampled, and generated datasets. In all cases, the average test/reference ratio is 1 and the within-subject variability equals 20% (A), 40% (B), and 60% (C). Three different original sample sizes are used (24, 48, and 72), while the sizes of the generated 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 53 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 54. 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 54A) and CV = 40% (Figure 54B). It is important to underline that for the scenarios presented in Figure 54, the acceptance limits for the VAE-generated datasets were not extended, in contrast with the acceptance limits for the original and subsampled cases.

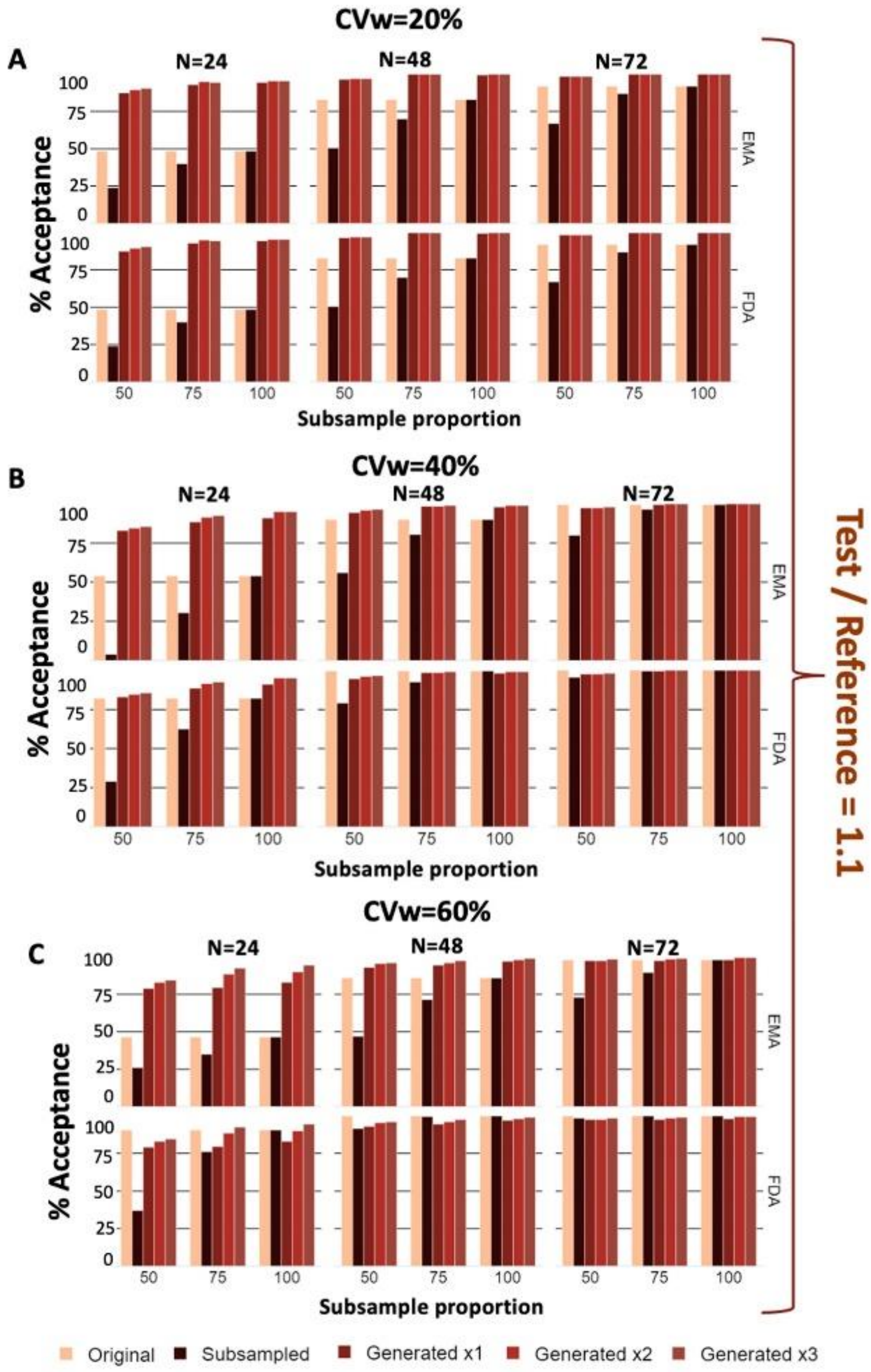


Figure 54. Acceptance rates of the original, subsampled, and generated datasets. In all cases, the average test/reference ratio is 1:1 (i.e., 10% mean difference) and the within-subject variability equals 20% (A), 40% (B), and 60% (C). Three different original sample sizes are used (24, 48, and 72), 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 9 and are presented in Figure 55 and Figure 56. In Figure 55, the ratio of the means of the T and R pharmaceutical products is equal to 1, while in Figure 56 the T/R ratio is 1:1 The scenarios that are not shown in Figure 55 and Figure 56 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.

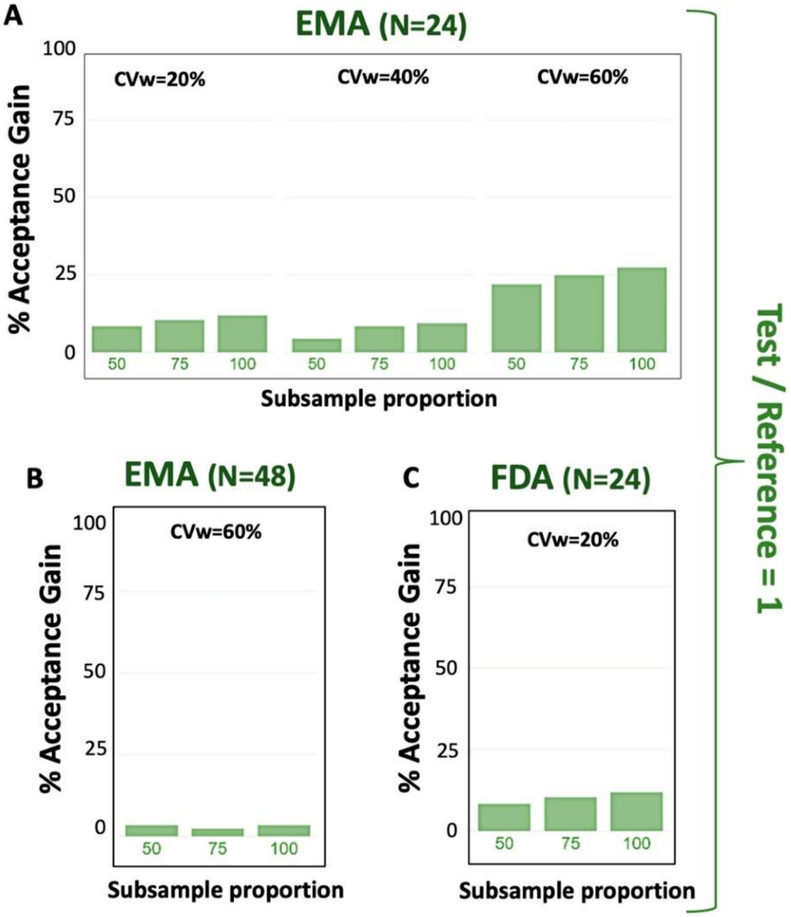


Figure 55. Acceptance gains, namely, the difference in the acceptance rate percentages between the VAE-generated and the original datasets. The acceptance limits for the original datasets are increased according to the EMA (A,B) and FDA (C) guidelines. In all cases, the average test/reference ratio is 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×) to that of the original sample. The scenarios where the generated and original datasets perform identically are not shown due to space restrictions.

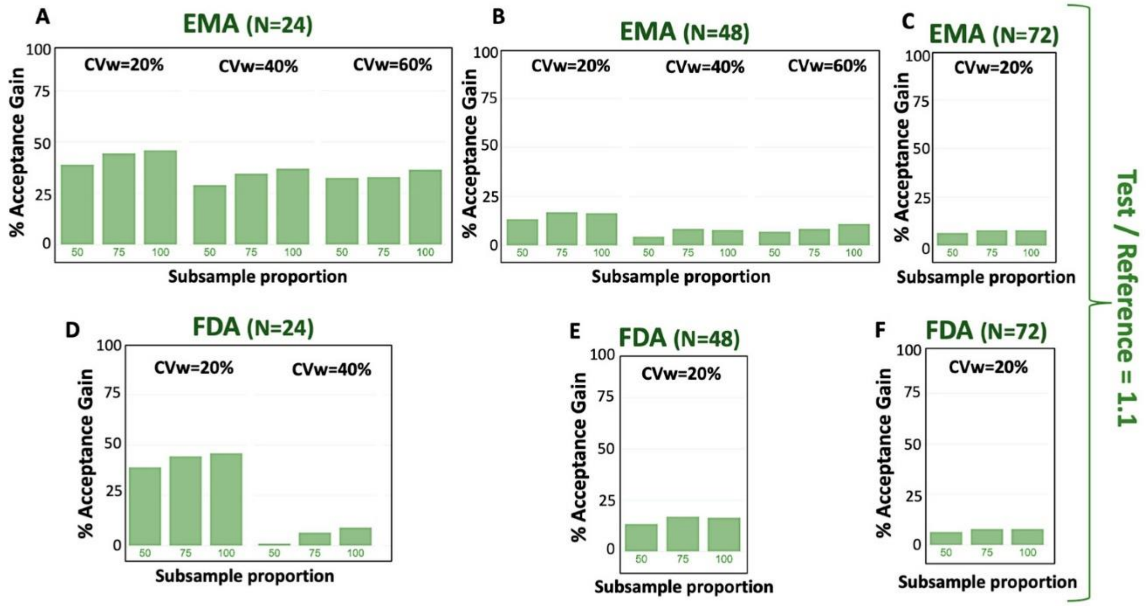


Figure 56. Acceptance gains, namely, the difference in the acceptance rate percentages between the VAE-generated and the original datasets. The acceptance limits for the original datasets are increased according to the EMA (A–C) and FDA (D–F) guidelines. In all cases, the average test/reference ratio 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×) to that of the original sample. The scenarios where the generated and original datasets perform identically are not shown due to space restrictions.

Figure 55 illustrates that the acceptance gain for the EMA scenarios (Figure 55A and 59B 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 55C), 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 56 presents the acceptance gain for the scenario where T/R equals 1:1. Figure 56A, 60B and 60C show the acceptance gain when the limits are extended according to the EMA for sample sizes of 24, 48, and 72, respectively. Figure 56D–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 56A–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 56D–F) is similar to that of the EMA, though there are more scenarios where no acceptance gain was observed.

Figure 57 and Figure 58 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.

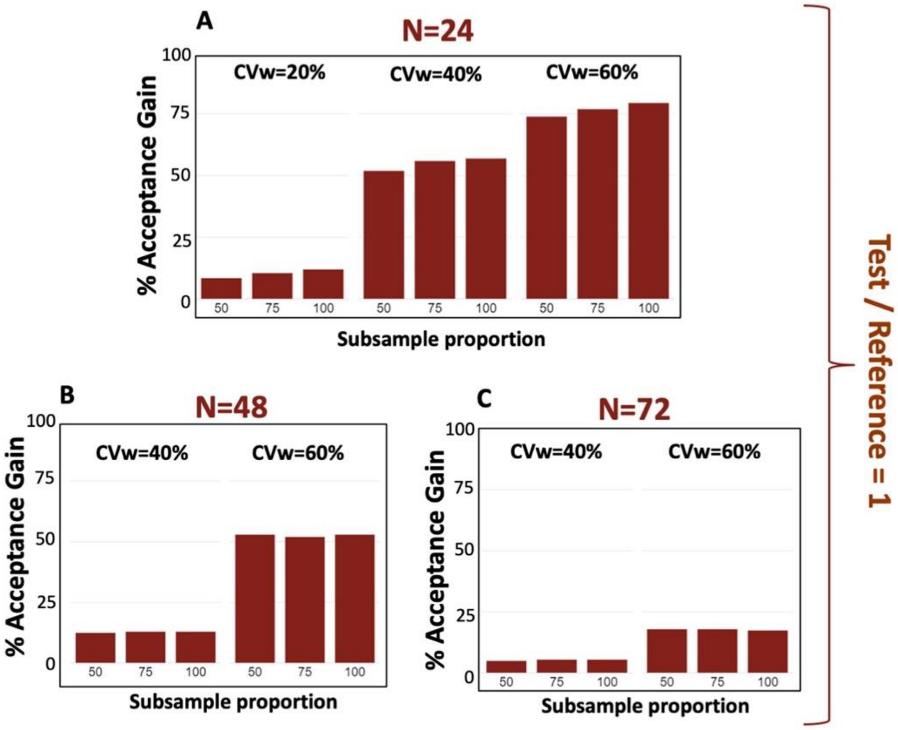


Figure 57. Acceptance gains between the VAE-generated and the original datasets, when no scaling is applied to either dataset. Three different original sample sizes are used (24 (A), 48 (B), and 72 (C)), while the size of the generated datasets is equal (1×) to that of the original sample. In all cases, the average test/reference ratio is 1 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.

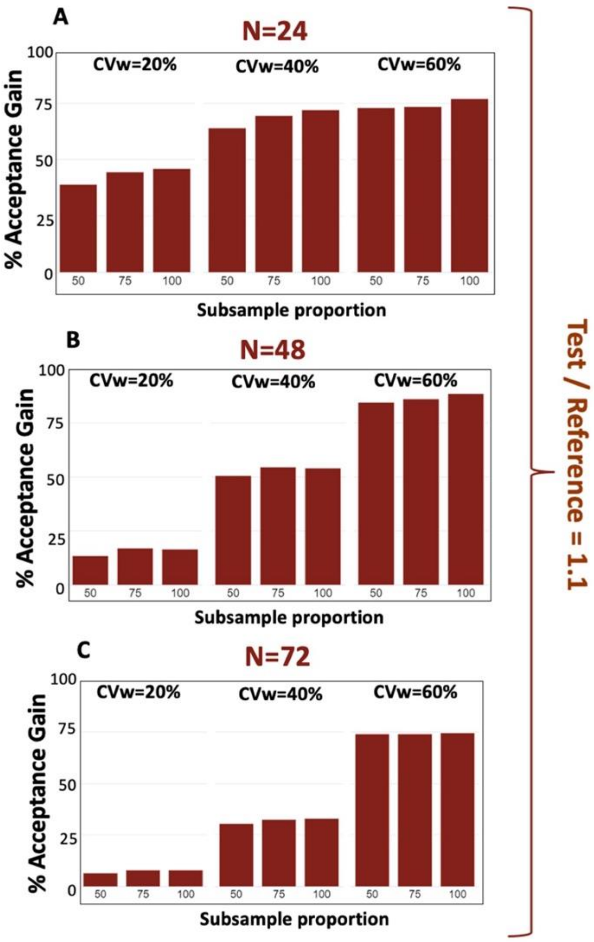


Figure 58. Acceptance gains between the VAE-generated and the original datasets when no scaling is applied to either dataset. Three different original sample sizes are used (24 (A), 48 (B), and 72 (C)), while the size of the generated datasets is equal (1×) to that of the original sample. In all cases, the average test/reference ratio is 1.1 and the within-subject variability equals 20%, 40%, and 60%.

From Figure 57, 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 CVW = 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 CVW increases and the sample size decreases.

Figure 58 is similar to Figure 57, 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 58C) and a 20% CVW, the acceptance rate goes up to 80% for a sample size equal to 48 and a CV of 70% (Figure

58B). The acceptance gain is high overall (i.e., more than 30%) in the case where the sample size is 24 (Figure 58A).

C4.4. 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 [56,62,63]. 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.

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 [56]. 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 [62,63]. 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 (Figure 53 and Figure 55). 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 CVW = 60%. Under the assumption that T/R equals 1:1, and CVw 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.

C4.5. 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.

Chapter D: Discussion

This thesis was designed with the purpose of applying machine learning and neural networks, which are the state-of-the-art models of deep learning, to improve the processes and results of clinical trials as well as bioequivalence studies. The goal was to apply these advanced computational techniques to gain deeper insights and enhance the accuracy and efficiency of the analyses conducted in these important areas of medical research.

The first goal of this dissertation was to explore and apply machine learning methods in bioequivalence and clinical studies. In this context, firstly the kinetics of NAbs and anti-S-RBDs against SARS-CoV-2 in 309 healthy individuals after receiving the BNT162b2 mRNA vaccine over a 9-month period were analyzed. It is shown that BNT162b2 provides decreasing but significant COVID-19 protection 6 and 9 months after full vaccination [97, 101, 112, 113, 114]. Though, results showed a gradual decline in both NAbs and anti-S-RBD IgGs. Simulation studies projected that median NAb levels would decrease from 66% at 9 months to 59% and 45% at 12 and 18 months post-vaccination, respectively. This suggests an immune protection against COVID-19 which is weakening, highlighting the need for booster shots, particularly in regions facing new outbreaks. Moreover, the identification of individual factors that predict the Nabs levels post-vaccination was performed. Data from 302 subjects were analyzed using machine learning techniques, including principal component analysis, factor analysis of mixed data, k-means clustering, and random forest. PCA and FAMD showed that younger individuals had higher NAbs levels compared to older ones, with the age effect being strongest near the vaccination date and diminishing over time. Obesity was linked to a lower antibody response [111]. Gender did not affect NAbs at nine months, though there was a slight association at earlier stages. Participants with autoimmune diseases had lower inhibitory levels than those without. K-means clustering identified five natural subject groups with predominant characteristics. Random forest ranked the importance of these characteristics, showing that older age, higher body mass index, and autoimmune diseases negatively impacted NAbs development nine months after full vaccination [116, 117, 118, 119].

Nevertheless, the basic idea of this thesis was to introduce the concept of data augmentation in clinical studies with the use of generative AI algorithms. So, the novel idea of using deep learning to virtually increase the sample size in clinical and bioequivalence studies was introduced. An original framework was created to simulate the exact conditions of these studies and neural networks were utilized to achieve this task.

Sample size estimation is a critical aspect of clinical trials and bioequivalence studies, as it ensures that the collected data can provide meaningful insights into the population under study [53, 130].

However, gathering large datasets is both costly and time-intensive. For this purpose, a new approach to data augmentation in clinical trials was introduced, using variational autoencoders. By developing and using various forms of VAEs, we were able to generate virtual subjects that closely mimic real data. The VAE-generated data demonstrated similar performance to the original data, even when only 30–40% of the original data was used for reconstruction. It is worth noting that, in scenarios with high variability, the VAE-generated data exhibited higher statistical power, effectively reducing noise and enhancing the robustness of the results. This suggests that VAEs can be a valuable tool in clinical trials, potentially decreasing the required sample size and thereby reducing costs and time, while also addressing ethical concerns related to human participation.

Building on this concept, we explored the application of VAEs in bioequivalence studies. A computational methodology was developed that combines Monte Carlo simulations of 2 × 2 crossover BE trials with deep learning algorithms, specifically VAEs. Various scenarios were examined, including different levels of variability, actual sample sizes, VAE-generated sample sizes, and performance differences between the pharmaceutical products under comparison. The simulations revealed that incorporating AI generative algorithms to create virtual populations in BE trials offers numerous advantages. It allows for the use of less actual human data while achieving similar or even superior results. This approach can significantly reduce human exposure, costs, and the time required to complete trials.

The study also addressed the bioequivalence for highly variable drugs, which is a significant challenge in the field of BE assessment [56, 57, 58] and can significantly impact the required sample size and statistical power. Regulatory agencies like the EMA and FDA recommend using scaled limits to manage this variability [56, 132]. We propose using VAEs to virtually increase sample size, thereby reducing the need for actual human subjects in BE studies of highly variable drugs. Monte Carlo simulations incorporating two levels of stochasticity (between-subject and within-subject) were used to create virtual populations. The performance of VAE-generated datasets was compared to traditional methods using constant 80–125% limits or scaled BE limits. The results showed that VAE-generated datasets outperformed both scaled and unscaled BE approaches, even with less than half the typically required sample size. This demonstrates the potential of VAEs to enhance the efficiency of BE studies, aligning with ethical considerations and regulatory guidelines.

Overall, the study demonstrated that machine learning is incredibly powerful for analyzing and uncovering patterns in data in the field of clinical trials and supports the idea of leveraging AI in healthcare [7, 133, 135, 136, 137, 138]. By using algorithms that can learn from and make predictions based on data, ML enables the identification of complex patterns and trends that are difficult to detect otherwise. This capability is particularly useful where it is important to understand subtle data patterns and improve decision-making. Additionally, ML can handle vast amounts of data at high speeds, providing insights in real-time and allowing for more dynamic and responsive systems.

More importantly, this study explored for the first time, the use of VAEs for data augmentation in clinical and bioequivalence studies. It has shown that the application of VAEs in clinical and BE studies represents a modern tool that can significantly reduce the need for large numbers of human subjects, lower costs, and shorten trial completion times, while maintaining or even improving the quality and reliability of the data.

It is important, in the future, to test the VAE algorithm on actual clinical data. Integrating real data into simulations is crucial for improving their accuracy, reliability, and relevance, making them valuable for understanding real-world phenomena, making predictions, and supporting decision-making. A comparative analysis of the VAE procedure applied to real clinical data would help assess its effectiveness and suitability, since although simulated data are consistent and reproducible [134, 138, 139, 140, 141], they often lack the intricacies and subtleties found in real-world data. Real data provide a more accurate and trustworthy basis for analysis, ensuring that the methods proposed are practical and effective in real-life applications. Moreover, real data can expose unexpected challenges and variables that simulations might miss, leading to more robust and widely applicable conclusions. Therefore, incorporating real data is essential for transitioning from theoretical research to practical implementation, ensuring the relevance and impact of the findings.

Chapter E: Concluding remarks

This study focused on the application of machine learning algorithms and deep learning models in clinical trials and bioequivalence studies. In this context, a novel framework was created, which combined Monte Carlo simulations with a generative neural network model, namely variational autoencoders to virtually increase the sample size of clinical studies.

It is evident that machine learning has strong analytical power that can help discover relationships and patterns in the clinical trial domain, which would be impossible to achieve with the use of classical statistical approaches. In this case, grouping the individuals from a clinical trial into five distinct groups and identifying the parameters responsible for NAbs levels nine months after BNT162b2 vaccination.

Generative neural networks like variational autoencoders were utilized to virtually increase the sample size of clinical trials and bioequivalence studies. Firstly, the most appropriate architecture for the VAE was investigated by tuning the respective hyperparameters. Later the optimized neural networked was used to create synthetic data, based on real data under different simulated scenarios. These scenarios included various levels of variability in the measured endpoints, differing average performances between the comparison groups, a range of subsampled and actual group sizes between the T and R products, in the context of clinical trials and bioequivalence studies.

The simulations carried out in this study highlighted several advantages of incorporating AI generative algorithms into clinical trials and bioequivalence studies for creating virtual populations. These advantages include reduced human exposure, significantly shorter study durations, simplified trial processes, decreased workload for physicians and clinics, and notably lower costs for sponsors or health agencies. The results showed that less actual human data can be used to achieve similar or even better outcomes in terms of statistical power. Overall, this study supports the use of AI-driven generative algorithms in clinical research.

Bibliography

1. Turing, A. M. Computing Machinery and Intelligence. In Parsing the Turing Test; Mind, 2009; pp. 23–65. https://doi.org/10.1007/978-1-4020-6710-5_3.
2. Zhang, C.; Lu, Y. Study on Artificial Intelligence: The State of the Art and Future Prospects. Journal of Industrial Information Integration 2021, 23 (23), 100224. https://doi.org/10.1016/j.jii.2021.100224.
3. Boute, R. N.; Udenio, M. AI in Logistics and Supply Chain Management. Global Logistics and Supply Chain Strategies for the 2020s 2022, 49–65. https://doi.org/10.1007/978-3-030-95764-3_3.
4. Chung, S.-H. Applications of Smart Technologies in Logistics and Transport: A Review. Transportation Research Part E: Logistics and Transportation Review 2021, 153 (1), 102455. https://doi.org/10.1016/j.tre.2021.102455.
5. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence 2019, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007.
6. Pomerol, J.-C. Artificial Intelligence and Human Decision Making. European Journal of Operational Research 1997, 99 (1), 3–25. https://doi.org/10.1016/s0377-2217(96)00378-5.
7. Jiang, F.; Jiang, Y.; Zhi, H. Artificial Intelligence in Healthcare: Past, Present and Future. Stroke and Vascular Neurology 2017, 2 (4), 230–243. https://doi.org/10.1136/svn-2017-000101.
8. Patel, V. L.; Shortliffe, E. H.; Stefanelli, M.; Szolovits, P.; Berthold, M. R.; Bellazzi, R.; Abu-Hanna, A. The Coming of Age of Artificial Intelligence in Medicine. Artificial Intelligence in Medicine 2009, 46 (1), 5–17. https://doi.org/10.1016/j.artmed.2008.07.017.
9. Yu, K.-H.; Beam, A. L.; Kohane, I. S. Artificial Intelligence in Healthcare. Nature Biomedical Engineering 2018, 2 (10), 719–731. https://doi.org/10.1038/s41551-018-0305-z.
10. Dilsizian, S. E.; Siegel, E. L. Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment. Current Cardiology Reports 2013, 16 (1). https://doi.org/10.1007/s11886-013-0441-8.

11. Jha, S.; Topol, E. J. Adapting to Artificial Intelligence. JAMA 2016, 316 (22), 2353. https://doi.org/10.1001/jama.2016.17438.

12. Shaheen, M. Y. Applications of Artificial Intelligence (AI) in Healthcare: A Review. ScienceOpen Preprints 2021, 1 (1). https://doi.org/10.14293/s2199-1006.1.sor-.ppvry8k.v1.

13. Chan, H. C. S.; Shan, H.; Dahoun, T.; Vogel, H.; Yuan, S. Advancing Drug Discovery via Artificial Intelligence. Trends in Pharmacological Sciences 2019, 40 (8), 592–604. https://doi.org/10.1016/j.tips.2019.06.004.

14. Díaz, Ó.; Dalton, J. A. R.; Giraldo, J. Artificial Intelligence: A Novel Approach for Drug Discovery. Trends in Pharmacological Sciences 2019, 40 (8), 550–551. https://doi.org/10.1016/j.tips.2019.06.005.

15. Tang, A.; Tam, R.; Cadrin-Chênevert, A.; Guest, W.; Chong, J.; Barfett, J.; Chepelev, L.; Cairns, R.; Mitchell, J. R.; Cicero, M. D.; Poudrette, M. G.; Jaremko, J. L.; Reinhold, C.; Gallix, B.; Gray, B.; Geis, R.; O'Connell, T.; Babyn, P.; Koff, D.; Ferguson, D. Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. Canadian Association of Radiologists Journal 2018, 69 (2), 120–135. https://doi.org/10.1016/j.carj.2018.02.002.

16. Safdar, N. M.; Banja, J. D.; Meltzer, C. C. Ethical Considerations in Artificial Intelligence. European Journal of Radiology 2020, 122 (1), 108768.

17. Beil, M.; Proft, I.; van Heerden, D.; Sviri, S.; van Heerden, P. V. Ethical Considerations about Artificial Intelligence for Prognostication in Intensive Care. Intensive Care Medicine Experimental 2019, 7 (1). https://doi.org/10.1186/s40635-019-0286-6.

18. Queensland Brain Institute. History of Artificial Intelligence. Uq.edu.au. https://qbi.uq.edu.au/brain/intelligent-machines/history-artificial-intelligence.

19. Jordan, M. I.; Mitchell, T. M. Machine Learning: Trends, Perspectives, and Prospects. Science 2020, 349 (6245), 255–260. https://doi.org/10.1126/science.aaa8415.

20. El Naqa, I.; Murphy, M. J. What Is Machine Learning? Machine Learning in Radiation Oncology 2015, 1 (1), 3–11. https://doi.org/10.1007/978-3-319-18305-3_1.

21. Zhi-Hua Zhou; Shaowu Liu. Machine Learning; Singapore Springer, 2021.

22. Ethem Alpaydin. Machine Learning : The New AI; Mit Press: Cambridge, Ma, 2016.

23. Kanade, V. What Is Machine Learning? Definition, Types, Applications, and Trends for 2022. Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/

24. Zebari, R.; Abdulazeez, A.; Zeebaree, D.; Zebari, D.; Saeed, J. A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends 2020, 1 (2), 56–70. https://doi.org/10.38094/jastt1224.

25. Hady, M. F. A.; Schwenker, F. Semi-Supervised Learning. Intelligent Systems Reference Library 2013, 215–239. https://doi.org/10.1007/978-3-642-36657-4_7.

26. Reinforcement Learning; Wiering, M., van Otterlo, M., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012. https://doi.org/10.1007/978-3-642-27645-3.

27. Cinarer, G.; Emiroglu, B. G. Classificatin of Brain Tumors by Machine Learning Algorithms. 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2019. https://doi.org/10.1109/ismsit.2019.8932878.

28. Erickson, B. J.; Korfiatis, P.; Akkus, Z.; Kline, T. L. Machine Learning for Medical Imaging. RadioGraphics 2017, 37 (2), 505–515. https://doi.org/10.1148/rg.2017160130.

29. Awoyemi, J. O.; Adetunmbi, A. O.; Oluwadare, S. A. Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI) 2017, 1–9. https://doi.org/10.1109/iccni.2017.8123782.

30. Perols, J. Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. AUDITING: A Journal of Practice & Theory 2011, 30 (2), 19–50. https://doi.org/10.2308/ajpt-50009.

31. Nuti, G.; Mirghaemi, M.; Treleaven, P.; Yingsaeree, C. Algorithmic Trading. Computer 2011, 44 (11), 61–69. https://doi.org/10.1109/mc.2011.31.

32. Dastile, X.; Celik, T.; Potsane, M. Statistical and Machine Learning Models in Credit Scoring: A Systematic Literature Survey. Applied Soft Computing 2020, 91, 106263. https://doi.org/10.1016/j.asoc.2020.106263.

33. Steck, H.; Baltrunas, L.; Elahi, E.; Liang, D.; Raimond, Y.; Basilico, J. Deep Learning for Recommender Systems: A Netflix Case Study. AI Magazine 2021, 42 (3), 7–18.

34. Jacobson, K.; Murali, V.; Newett, E.; Whitman, B.; Yon, R. Music Personalization at Spotify. Proceedings of the 10th ACM Conference on Recommender Systems - RecSys ’16 2016. https://doi.org/10.1145/2959100.2959120.

35. Stilgoe, J. Machine Learning, Social Learning and the Governance of Self-Driving Cars. Social Studies of Science 2017, 48 (1), 25–56. https://doi.org/10.1177/0306312717741687.

36. Klare, B. F.; Burge, M. J.; Klontz, J. C.; Vorder Bruegge, R. W.; Jain, A. K. Face Recognition Performance: Role of Demographic Information. IEEE Transactions on Information Forensics and Security 2012, 7 (6), 1789–1801. https://doi.org/10.1109/tifs.2012.2214212.

37. White, D.; Dunn, J. D.; Schmid, A. C.; Kemp, R. I. Error Rates in Users of Automatic Face Recognition Software. PLOS ONE 2015, 10 (10), e0139827. https://doi.org/10.1371/journal.pone.0139827.

38. Banks, V. A.; Plant, K. L.; Stanton, N. A. Driver Error or Designer Error: Using the Perceptual Cycle Model to Explore the Circumstances Surrounding the Fatal Tesla Crash on 7th May 2016. Safety Science 2018, 108, 278–285. https://doi.org/10.1016/j.ssci.2017.12.023.

39. Angelov, P. P.; Soares, E. A.; Jiang, R.; Arnold, N. I.; Atkinson, P. M. Explainable Artificial Intelligence: An Analytical Review. WIREs Data Mining and Knowledge Discovery 2021, 11 (5). https://doi.org/10.1002/widm.1424.

40. Dosilovic, F. K.; Brcic, M.; Hlupic, N. Explainable Artificial Intelligence: A Survey. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) 2018. https://doi.org/10.23919/mipro.2018.8400040.

41. Minh, D.; Wang, H. X.; Li, Y. F.; Nguyen, T. N. Explainable Artificial Intelligence: A Comprehensive Review. Artificial Intelligence Review 2021, 55. https://doi.org/10.1007/s10462-021-10088-y.

42. Krogh, A. What Are Artificial Neural Networks? Nature Biotechnology 2008, 26 (2), 195–197. https://doi.org/10.1038/nbt1386.

43. Picton, P. What Is a Neural Network? Introduction to Neural Networks 1994, 1–12. https://doi.org/10.1007/978-1-349-13530-1_1.

44. Bank, D.; Noam Koenigstein; Raja Giryes. Autoencoders. Springer eBooks 2023, 353–374. https://doi.org/10.1007/978-3-031-24628-9_16.

45. Lopez Pinaya, W. H.; Vieira, S.; Garcia-Dias, R.; Mechelli, A. Chapter 11 - Autoencoders. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/B9780128157398000110.

46. Charte, D.; Charte, F.; del Jesus, M. J.; Herrera, F. An Analysis on the Use of Autoencoders for Representation Learning: Fundamentals, Learning Task Case Studies, Explainability and Challenges. Neurocomputing 2020, 404, 93–107. https://doi.org/10.1016/j.neucom.2020.04.057.

47. Islam, Z.; Abdel-Aty, M.; Cai, Q.; Yuan, J. Crash Data Augmentation Using Variational Autoencoder. Accident Analysis & Prevention 2021, 151, 105950. https://doi.org/10.1016/j.aap.2020.105950.

48. Papadopoulos, D.; Karalis, V. D. Variational Autoencoders for Data Augmentation in Clinical Studies. Applied Sciences 2023, 13 (15), 8793. https://doi.org/10.3390/app13158793.

49. Papadopoulos, D.; Karalis, V. D. Introducing an Artificial Neural Network for Virtually Increasing the Sample Size of Bioequivalence Studies. Applied Sciences 2024, 14 (7), 2970–2970. https://doi.org/10.3390/app14072970.

50. Papadopoulos, D.; Karali, G.; Karalis, V. D. Bioequivalence Studies of Highly Variable Drugs: An Old Problem Addressed by Artificial Neural Networks. Applied Sciences 2024, 14 (12), 5279–5279. https://doi.org/10.3390/app14125279.

51. Tanaka, S.; Claus Aranha. Data Augmentation Using GANs. arXiv (Cornell University) 2019. https://doi.org/10.48550/arxiv.1904.09135.

52. Chatziagapi, A.; Paraskevopoulos, G.; Sgouropoulos, D.; Pantazopoulos, G.; Nikandrou, M.; Giannakopoulos, T.; Katsamanis, A.; Potamianos, A.; Narayanan, S. Data Augmentation Using GANs for Speech Emotion Recognition. Interspeech 2019 2019. https://doi.org/10.21437/interspeech.2019-2561.

53. Sakpal, T. V. Sample Size Estimation in Clinical Trial. Perspectives in Clinical Research 2010, 1 (2), 67.

54. Wang, X.; Ji, X. Sample Size Estimation in Clinical Research: From Randomized Controlled Trials to Observational Studies. CHEST 2020, 158 (1), S12–S20. https://doi.org/10.1016/j.chest.2020.03.010.

55. Helen Evelyn Malone , H. N. Fundamentals of estimating sample size. journals.rcni.com. https://journals.rcni.com/nurse-researcher/fundamentals-of-estimating-sample-size-nr.23.5.21.s5.

56. Karalis, V. Modeling and Simulation in Bioequivalence. Interdisciplinary applied mathematics 2016, 227–254. https://doi.org/10.1007/978-3-319-27598-7_10.

57. 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).

58. 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-inds-general-considerations (accessed on 29 May 2023).

59. Askin, S.; Burkhalter, D.; Calado, G.; El Dakrouni, S. Artificial Intelligence Applied to Clinical Trials: Opportunities and Challenges. Health Technol. 2023, 13, 203–213.

60. Harrer, S.; Shah, P.; Antony, B.; Hu, J. Artificial Intelligence for Clinical Trial Design. Trends Pharmacol. Sci. 2019, 40, 577–591.

61. 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.

62. 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).

63. 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).

64. Andriy Burkov. THE HUNDRED-PAGE MACHINE LEARNING BOOK; Andriy Burkov, 2019.

65. Cross Validation, Explained - Sharp Sight. www.sharpsightlabs.com. https://www.sharpsightlabs.com/blog/cross-validation-explained/.

66. Christiaan Heij; Paul de Boer; Philip Hans Franses; Teun Kloek; Herman; Rotterdam, in. Econometric Methods with Applications in Business and Economics; OUP Oxford, 2004.

67. STHDA - Home. www.sthda.com. http://www.sthda.com/english/.

68. 1.10. Decision Trees. scikit-learn. https://scikit-learn.org/stable/modules/tree.html#mathematical-formulation.
69. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition; Springer New York: New York, Ny, 2009.
70. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 2016, 785–794.
71. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 2017, 30.
72. How XGBoost Works - Amazon SageMaker. docs.aws.amazon.com. https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html.
73. Pang-Ning Tan; Steinbach, M.; Vipin Kumar. Introduction to Data Mining; Pearson Education: San Francisco, 2006.
74. Zhang, F.; O'Donnell, L. J. Chapter 7 - Support vector regression. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/B9780128157398000079.
75. Rusdah, D. A.; Murfi, H. XGBoost in Handling Missing Values for Life Insurance Risk Prediction. SN Applied Sciences 2020, 2 (8). https://doi.org/10.1007/s42452-020-3128-y.
76. Shai Shalev-Shwartz; Shai Ben-David. Understanding Machine Learning : From Foundations to Algorithms; Cambridge University Press: Cambridge Etc, 2014.
77. Piech, C. CS221. Stanford.edu. https://stanford.edu/~cpiech/cs221/handouts/kmeans.html.
78. Baig, A. M.; Ardakani, E. P. Using Machine Learning to Estimate the Flow of Stress Using Microseismicity Recorded during Hydraulic Fracturing. 2018. https://doi.org/10.1190/segam2018-2992584.1.
79. Bishop, C. M. Neural Networks for Pattern Recognition; Oxford University Press: Oxford, 2002.
80. Pramoditha, R. Overview of a Neural Network's Learning Process. Data Science 365. https://medium.com/data-science-365/overview-of-a-neural-networks-learning-process-61690a502fa.
81. Irekponor, V. E. Mathematical Prerequisites For Understanding Autoencoders and Variational Autoencoders (VAEs). Analytics Vidhya. https://medium.com/analytics-

vidhya/mathematical-prerequisites-for-understanding-autoencoders-and-variational-autoencoders-vaes-8f854025390e.

82. Doersch, C. Tutorial on Variational Autoencoders. arXiv:1606.05908 [cs, stat] 2016.
83. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. Communications of the ACM 2020, 63 (11), 139–144. https://doi.org/10.1145/3422622.
84. Tsang, H. F.; Chan, L. W. C.; Cho, W. C. S.; Yu, A. C. S.; Yim, A. K. Y.; Chan, A. K. C.; Ng, L. P. W.; Wong, Y. K. E.; Pei, X. M.; Li, M. J. W.; Wong, S. C. C. An Update on COVID-19 Pandemic: The Epidemiology, Pathogenesis, Prevention and Treatment Strategies. Expert Review of Anti-infective Therapy 2020, 19 (7). https://doi.org/10.1080/14787210.2021.1863146.
85. Tentolouris, A.; Ntanasis-Stathopoulos, I.; Vlachakis, P. K.; Tsilimigras, D. I.; Gavriatopoulou, M.; Dimopoulos, M. A. COVID-19: Time to Flatten the Infodemic Curve. Clinical and Experimental Medicine 2021. https://doi.org/10.1007/s10238-020-00680-x.
86. Beyerstedt, S.; Casaro, E. B.; Rangel, É. B. COVID-19: Angiotensin-Converting Enzyme 2 (ACE2) Expression and Tissue Susceptibility to SARS-CoV-2 Infection. European Journal of Clinical Microbiology & Infectious Diseases 2021, 40 (5). https://doi.org/10.1007/s10096-020-04138-6.
87. Korompoki, E.; Gavriatopoulou, M.; Fotiou, D.; Ntanasis-Stathopoulos, I.; Dimopoulos, M. A.; Terpos, E. Late Onset Hematological Complications Post COVID -19: An Emerging Medical Problem for the Hematologist. American Journal of Hematology 2021. https://doi.org/10.1002/ajh.26384.
88. Gavriatopoulou, M.; Korompoki, E.; Fotiou, D.; Ntanasis-Stathopoulos, I.; Psaltopoulou, T.; Kastritis, E.; Terpos, E.; Dimopoulos, M. A. Organ-Specific Manifestations of COVID-19 Infection. Clinical and Experimental Medicine 2020, 20 (4), 493–506. https://doi.org/10.1007/s10238-020-00648-x.
89. Korompoki, E.; Gavriatopoulou, M.; Hicklen, R. S.; Ntanasis-Stathopoulos, I.; Kastritis, E.; Fotiou, D.; Stamatelopoulos, K.; Terpos, E.; Kotanidou, A.; Hagberg, C. A.; Dimopoulos, M. A.; Kontoyiannis, D. P. Epidemiology and Organ Specific Sequelae of Post-Acute COVID19: A Narrative Review. The Journal of Infection 2021. https://doi.org/10.1016/j.jinf.2021.05.004.
90. The Lancet Infectious Diseases. COVID-19 Vaccine Equity and Booster Doses. The Lancet Infectious Diseases 2021, 21 (9), 1193. https://doi.org/10.1016/s1473-3099(21)00486-2.

91. Haas, E. J.; Angulo, F. J.; McLaughlin, J. M.; Anis, E.; Singer, S. R.; Khan, F.; Brooks, N.; Smaja, M.; Mircus, G.; Pan, K.; Southern, J.; Swerdlow, D. L.; Jodar, L.; Levy, Y.; Alroy-Preis, S. Impact and Effectiveness of MRNA BNT162b2 Vaccine against SARS-CoV-2 Infections and COVID-19 Cases, Hospitalisations, and Deaths Following a Nationwide Vaccination Campaign in Israel: An Observational Study Using National Surveillance Data. The Lancet 2021, 0 (0). https://doi.org/10.1016/S0140-6736(21)00947-8.

92. Polack, F. P.; Thomas, S. J.; Kitchin, N.; Absalon, J.; Gurtman, A.; Lockhart, S.; Perez, J. L.; Pérez Marc, G.; Moreira, E. D.; Zerbini, C.; Bailey, R.; Swanson, K. A.; Roychoudhury, S.; Koury, K.; Li, P.; Kalina, W. V.; Cooper, D.; Frenck, R. W.; Hammitt, L. L.; Türeci, Ö. Safety and Efficacy of the BNT162b2 MRNA Covid-19 Vaccine. New England Journal of Medicine 2020, 383 (27), 2603–2615. https://doi.org/10.1056/nejmoa2034577.

93. Jalkanen, P.; Kolehmainen, P.; Häkkinen, H. K.; Huttunen, M.; Tähtinen, P. A.; Lundberg, R.; Maljanen, S.; Reinholm, A.; Tauriainen, S.; Pakkanen, S. H.; Levonen, I.; Nousiainen, A.; Miller, T.; Välimaa, H.; Ivaska, L.; Pasternack, A.; Naves, R.; Ritvos, O.; Österlund, P.; Kuivanen, S. COVID-19 MRNA Vaccine Induced Antibody Responses against Three SARS-CoV-2 Variants. Nature Communications 2021, 12. https://doi.org/10.1038/s41467-021-24285-4.

94. Rosati, M.; Evangelos Terpos; Agarwal, M.; Karalis, V.; Bear, J.; Burns, R.; Hu, X.; Demetrios Papademetriou; Ioannis Ntanasis-Stathopoulos; Trougakos, I. P.; Dimopoulos, M.; Pavlakis, G. N.; Felber, B. K. Distinct Neutralization Profile of Spike Variants by Antibodies Induced upon SARS-CoV-2 Infection or Vaccination. American Journal of Hematology 2021, 97 (1). https://doi.org/10.1002/ajh.26380.

95. Khoury, D. S.; Cromer, D.; Reynaldi, A.; Schlub, T. E.; Wheatley, A. K.; Juno, J. A.; Subbarao, K.; Kent, S. J.; Triccas, J. A.; Davenport, M. P. Neutralizing Antibody Levels Are Highly Predictive of Immune Protection from Symptomatic SARS-CoV-2 Infection. Nature Medicine 2021, 27 (27), 1–7. https://doi.org/10.1038/s41591-021-01377-8.

96. Garcia-Beltran, W. F.; Lam, E. C.; Astudillo, M. G.; Yang, D.; Miller, T. E.; Feldman, J.; Hauser, B. M.; Caradonna, T. M.; Clayton, K. L.; Nitido, A. D.; Murali, M. R.; Alter, G.; Charles, R. C.; Dighe, A.; Branda, J. A.; Lennerz, J. K.; Lingwood, D.; Schmidt, A. G.; Iafrate, A. J.; Balazs, A. B. COVID-19-Neutralizing Antibodies Predict Disease Severity and Survival. Cell 2021, 184 (2), 476-488.e11. https://doi.org/10.1016/j.cell.2020.12.015.

97. Wall, E. C.; Wu, M.; Harvey, R.; Kelly, G.; Warchal, S.; Sawyer, C.; Daniels, R.; Hobson, P.; Hatipoglu, E.; Ngai, Y.; Hussain, S.; Nicod, J.; Goldstone, R.; Ambrose, K.; Hindmarsh, S.; Beale, R.; Riddell, A.; Gamblin, S.; Howell, M.; Kassiotis, G. Neutralising Antibody Activity against SARS-CoV-2 VOCs B.1.617.2 and B.1.351 by BNT162b2 Vaccination. The Lancet 2021, 0 (0). https://doi.org/10.1016/S0140-6736(21)01290-3.

98. Terpos, E.; Trougakos, I. P.; Karalis, V.; Ntanasis-Stathopoulos, I.; Gumeni, S.; Apostolakou, F.; Sklirou, A. D.; Gavriatopoulou, M.; Skourti, S.; Kastritis, E.; Korompoki, E.; Papassotiriou, I.; Dimopoulos, M. A. Kinetics of Anti-SARS-CoV-2 Antibody Responses 3 Months Post Complete Vaccination with BNT162b2; a Prospective Study in 283 Health Workers. Cells 2021, 10 (8), 1942. https://doi.org/10.3390/cells10081942.

99. Goldberg, Y.; Mandel, M.; Bar-On, Y. M.; Bodenheimer, O.; Freedman, L.; Haas, E. J.; Milo, R.; Alroy-Preis, S.; Ash, N.; Huppert, A. Waning Immunity after the BNT162b2 Vaccine in Israel. New England Journal of Medicine 2021, 385 (24). https://doi.org/10.1056/nejmoa2114228.

100. Terpos, E.; Karalis, V.; Ntanasis-Stathopoulos, I.; Gavriatopoulou, M.; Gumeni, S.; Malandrakis, P.; Papanagnou, E.-D.; Kastritis, E.; Trougakos, I. P.; Dimopoulos, M. A. Robust Neutralizing Antibody Responses 6 Months Post Vaccination with BNT162b2: A Prospective Study in 308 Healthy Individuals. Life 2021, 11 (10), 1077. https://doi.org/10.3390/life11101077.

101. Naaber, P.; Tserel, L.; Kangro, K.; Sepp, E.; Jurjenson, V.; Adamson, A.; Haljasmagi, L.; Rumm, A.P.; Maruste, R.; Karner, J.; et al. Dynamics of antibody response to BNT162b2 vaccine after six months: A longitudinal prospective study. Lancet Reg. Health Eur. 2021, 10, 100208https://doi.org/10.1016/j.lanepe.2021.100208.

102. Bar-On, Y. M.; Goldberg, Y.; Mandel, M.; Bodenheimer, O.; Freedman, L.; Kalkstein, N.; Mizrahi, B.; Alroy-Preis, S.; Ash, N.; Milo, R.; Huppert, A. Protection of BNT162b2 Vaccine Booster against Covid-19 in Israel. New England Journal of Medicine 2021, 385 (15). https://doi.org/10.1056/nejmoa2114255.

103. Rella, S. A.; Kulikova, Y. A.; Dermitzakis, E. T.; Kondrashov, F. A. Rates of SARS-CoV-2 Transmission and Vaccination Impact the Fate of Vaccine-Resistant Strains. Scientific Reports 2021, 11 (1), 15729. https://doi.org/10.1038/s41598-021-95025-3.

104. Agrati, C.; Castilletti, C.; Goletti, D.; Meschi, S.; Sacchi, A.; Matusali, G.; Bordoni, V.; Petrone, L.; Lapa, D.; Notari, S.; Vanini, V.; Colavita, F.; Aiello, A.; Agresta, A.; Farroni, C.;

Grassi, G.; Leone, S.; Vaia, F.; Capobianchi, M. R.; Ippolito, G. Coordinate Induction of Humoral and Spike Specific T-Cell Response in a Cohort of Italian Health Care Workers Receiving BNT162b2 MRNA Vaccine. Microorganisms 2021, 9 (6), 1315. https://doi.org/10.3390/microorganisms9061315.

105. Cassaniti I, Bergami F, Percivalle E, et al. Humoral and cell-mediated response elicited by SARS-CoV-2 mRNA vaccine BNT162b2 e in healthcare workers: a longitudinal observational study. Clin Microbiol Infect. 2021 September 25

106. Mateus, J.; Dan, J. M.; Zhang, Z.; Rydyznski Moderbacher, C.; Lammers, M.; Goodwin, B.; Sette, A.; Crotty, S.; Weiskopf, D. Low-Dose MRNA-1273 COVID-19 Vaccine Generates Durable Memory Enhanced by Cross-Reactive T Cells. Science 2021, 374 (6566). https://doi.org/10.1126/science.abj9853.

107. Terpos, E.; Karalis, V.; Ntanasis-Stathopoulos, I.; Apostolakou, F.; Gumeni, S.; Gavriatopoulou, M.; Papadopoulos, D.; Malandrakis, P.; Papanagnou, E.; Korompoki, E.; et al. Sustained but declining humoral immunity against SARS-CoV-2 at 9 months post vaccination with BNT162b2: A prospective evaluation in 309 healthy individuals. Hemasphere 2022, 6, e677

108. Camacho, D.M.; Collins, K.M.; Powers, R.K.; Costello, J.C.; Collins, J.J. Next-Generation Machine Learning for Biological Networks. Cell 2018, 173, 1581–1592.

109. Shamout, F.; Zhu, T.; Clifton, D.A. Machine Learning for Clinical Outcome Prediction. IEEE Rev. Biomed. Eng. 2021, 14, 116–126

110. James, G.; Hastie, T.; Tibshirani, R.; Witten, D. An Introduction to Statistical Learning with Applications in R, 7th ed.; Springer: Berlin/Heidelberg, Germany, 2017

111. Westheim, A.J.F.; Bitorina, A.V.; Theys, J.; Shiri-Sverdlov, R. COVID-19 infection, progression, and vaccination: Focus on obesity and related metabolic disturbances. Obes. Rev. 2021, 22, e13313.

112. Campo, F.; Venuti, A.; Pimpinelli, F.; Abril, E.; Blandino, G.; Conti, L.; De Virgilio, A.; De Marco, F.; Di Noia, V.; Di Domenico, E. G.; Di Martino, S.; Ensoli, F.; Giannarelli, D.; Mandoj, C.; Mazzola, F.; Moretto, S.; Petruzzi, G.; Petrone, F.; Pichi, B.; Pontone, M. Antibody Persistence 6 Months Post-Vaccination with BNT162b2 among Health Care Workers. Vaccines 2021, 9 (10), 1125. https://doi.org/10.3390/vaccines9101125.

113. Tartof, S. Y.; Slezak, J. M.; Fischer, H.; Hong, V.; Ackerson, B. K.; Ranasinghe, O. N.; Frankland, T. B.; Ogun, O. A.; Zamparo, J. M.; Gray, S.; Valluri, S. R.; Pan, K.; Angulo, F. J.;

Jodar, L.; McLaughlin, J. M. Effectiveness of MRNA BNT162b2 COVID-19 Vaccine up to 6 Months in a Large Integrated Health System in the USA: A Retrospective Cohort Study. The Lancet 2021, 0 (0). https://doi.org/10.1016/S0140-6736(21)02183-8.

114. Thomas, S. J.; Moreira, E. D.; Kitchin, N.; Absalon, J.; Gurtman, A.; Lockhart, S.; Perez, J. L.; Pérez Marc, G.; Polack, F. P.; Zerbini, C.; Bailey, R.; Swanson, K. A.; Xu, X.; Roychoudhury, S.; Koury, K.; Bouguermouh, S.; Kalina, W. V.; Cooper, D.; Frenck, R. W.; Hammitt, L. L. Safety and Efficacy of the BNT162b2 MRNA Covid-19 Vaccine through 6 Months. New England Journal of Medicine 2021, 385 (19), 1761–1773. https://doi.org/10.1056/nejmoa2110345.

115. Terpos, E.; Trougakos, I.P.; Apostolakou, F.; Charitaki, I.; Sklirou, A.D.; Mavrianou, N.; Papanagnou, E.D.; Liacos, C.I.; Gumeni, S.; Rentziou, G.; et al. Age-dependent and gender-dependent antibody responses against SARS-CoV-2 in health workers and octogenarians after vaccination with the BNT162b2 mRNA vaccine. Am. J. Hematol. 2021, 96, E257–E259

116. Terpos, E.; Stellas, D.; Rosati, M.; Sergentanis, T. N.; Hu, X.; Politou, M.; Pappa, V.; Ntanasis-Stathopoulos, I.; Karaliota, S.; Bear, J.; Donohue, D.; Pagoni, M.; Grouzi, E.; Korompoki, E.; Pavlakis, G. N.; Felber, B. K.; Dimopoulos, M. A. SARS-CoV-2 Antibody Kinetics Eight Months from COVID-19 Onset: Persistence of Spike Antibodies but Loss of Neutralizing Antibodies in 24% of Convalescent Plasma Donors. European Journal of Internal Medicine 2021, 89, 87–96. https://doi.org/10.1016/j.ejim.2021.05.010.

117. Collier, D. A.; Ferreira, I. A. T. M.; Kotagiri, P.; Datir, R. P.; Lim, E. Y.; Touizer, E.; Meng, B.; Abdullahi, A.; Elmer, A.; Kingston, N.; Graves, B.; Le Gresley, E.; Caputo, D.; Bergamaschi, L.; Smith, K. G. C.; Bradley, J. R.; Ceron-Gutierrez, L.; Cortes-Acevedo, P.; Barcenas-Morales, G.; Linterman, M. A. Age-Related Immune Response Heterogeneity to SARS-CoV-2 Vaccine BNT162b2. Nature 2021, 596 (7872), 417–422. https://doi.org/10.1038/s41586-021-03739-1.

118. Muller L, Andree M, Moskorz W, et al. Age-dependent immune response to the Biontech/Pfizer BNT162b2 COVID-19 vaccination. Clin Infect Dis. 2021;73:2065–2072.

119. Terpos, E.; Trougakos, I.P.; Apostolakou, F.; Charitaki, I.; Sklirou, A.D.; Mavrianou, N.; Papanagnou, E.D.; Liacos, C.I.; Gumeni, S.; Rentziou, G.; et al. Age-dependent and gender-dependent antibody responses against SARS-CoV-2 in health workers and octogenarians after vaccination with the BNT162b2 mRNA vaccine. Am. J. Hematol. 2021, 96, E257–E259

120. Tober-Lau, P.; Schwarz, T.; Vanshylla, K.; Hillus, D.; Gruell, H.; Suttorp, N.; Landgraf, I.; Kappert, K.; Seybold, J.; Drosten, C.; Klein, F.; Kurth, F.; Sander, L. E.; Corman, V. M. Long-Term Immunogenicity of BNT162b2 Vaccination in Older People and Younger Health-Care Workers. The Lancet Respiratory Medicine 2021, 9 (11), e104–e105. https://doi.org/10.1016/S2213-2600(21)00456-2.

121. Kroon, F.P.B.; Najm, A.; Alunno, A.; Schoones, J.W.; Landewe, R.B.M.; Machado, P.M.; Navarro-Compan, V. Risk and prognosis of SARS-CoV-2 infection and vaccination against SARS-CoV-2 in rheumatic and musculoskeletal diseases: A systematic literature review to inform EULAR recommendations. Ann. Rheum. Dis. 2021.

122. Terpos, E.; Gavriatopoulou, M.; Fotiou, D.; Giatra, C.; Asimakopoulos, I.; Dimou, M.; Sklirou, A. D.; Ntanasis-Stathopoulos, I.; Darmani, I.; Briasoulis, A.; Kastritis, E.; Angelopoulou, M.; Baltadakis, I.; Panayiotidis, P.; Trougakos, I. P.; Vassilakopoulos, T. P.; Pagoni, M.; Dimopoulos, M. A. Poor Neutralizing Antibody Responses in 132 Patients with CLL, NHL and HL after Vaccination against SARS-CoV-2: A Prospective Study. Cancers 2021, 13 (17), 4480. https://doi.org/10.3390/cancers13174480.

123. Terpos, E.; Gavriatopoulou, M.; Ntanasis-Stathopoulos, I.; Briasoulis, A.; Gumeni, S.; Malandrakis, P.; Fotiou, D.; Papanagnou, E.-D.; Migkou, M.; Theodorakakou, F.; Roussou, M.; Eleutherakis-Papaiakovou, E.; Kanellias, N.; Trougakos, I. P.; Kastritis, E.; Dimopoulos, M. A. The Neutralizing Antibody Response Post COVID-19 Vaccination in Patients with Myeloma Is Highly Dependent on the Type of Anti-Myeloma Treatment. Blood Cancer Journal 2021, 11 (8). https://doi.org/10.1038/s41408-021-00530-3.

124. Gavriatopoulou, M.; Terpos, E.; Ntanasis-Stathopoulos, I.; Briasoulis, A.; Gumeni, S.; Malandrakis, P.; Fotiou, D.; Migkou, M.; Theodorakakou, F.; Eleutherakis-Papaiakovou, E.; et al. Poor neutralizing antibody responses in 106 patients with WM after vaccination against SARS-CoV-2; a prospective study. Blood Adv. 2021, 5, 4398–4405.

125. Gavriatopoulou, M.; Terpos, E.; Kastritis, E.; Briasoulis, A.; Gumeni, S.; Ntanasis-Stathopoulos, I.; Sklirou, A.D.; Malandrakis, P.; Eleutherakis-Papaiakovou, E.; Migkou, M.; et al. Low neutralizing antibody responses in WM, CLL and NHL patients after the first dose of the BNT162b2 and AZD1222 vaccine. Clin. Exp. Med. 2021, 1–5.

126. Corti, C.; Antonarelli, G.; Scotté, F.; Spano, J.; Barriére, J.; Michot, J.; André, F.; Curigliano, G. Seroconversion rate after vaccination against COVID-19 in cancer patients—A systematic review. Ann. Oncol. 2021.

127. 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

128. Kingma, D.; Welling, M. An Introduction to Variational Autoencoders (Foundations and Trends(r) in Machine Learning); Now Publishers Inc.: Hanover, MA, USA, 2019

129. Chollet, F. Deep Learning with Python, 2nd ed.; Manning; Simon and Schuster: New York, NY, USA, 2021.

130. Gupta, K.K.; Attri, J.P.; Singh, A.; Kaur, H.; Kaur, G. Basic Concepts for Sample Size Calculation: Critical Step for Any Clinical Trials. Saudi J. Anaesth. 2016, 10, 328–331

131. Lim, C.-Y. Considerations for Crossover Design in Clinical Study. Korean J. Anesthesiol. 2021, 74, 293–299

132. Endrenyi, L.; Tothfalusi, L. Bioequivalence for Highly Variable Drugs: Regulatory Agreements, Disagreements, and Harmonization. J. Pharmacokinet. Pharmacodyn. 2019, 46, 117–126.

133. Polevikov, S. Advancing AI in Healthcare: A Comprehensive Review of Best Practices. Clin. Chim. Acta 2023, 548, 117519

134. Ossowska, A.; Kusiak, A.; Świetlik, D. Artificial Intelligence in Dentistry—Narrative Review. International Journal of Environmental Research and Public Health 2022, 19 (6), 3449. https://doi.org/10.3390/ijerph19063449.

135. Karalis, V. D. The Integration of Artificial Intelligence into Clinical Practice. Applied Biosciences 2024, 3 (1), 14–44. https://doi.org/10.3390/applbiosci3010002.

136. Koski, E.; Murphy, J. AI in Healthcare. Studies in Health Technology and Informatics 2021, 284, 295–299. https://doi.org/10.3233/SHTI210726.

137. Chen, R. J.; Lu, M. Y.; Chen, T. Y.; Williamson, D. F. K.; Mahmood, F. Synthetic Data in Machine Learning for Medicine and Healthcare. Nature Biomedical Engineering 2021, 5 (6), 493–497. https://doi.org/10.1038/s41551-021-00751-8.

138. Ahsanullah Yunas Mahmoud; Neagu, D.; Scrimieri, D.; Abdullatif, A. Early Diagnosis and Personalised Treatment Focusing on Synthetic Data Modelling: Novel Visual Learning Approach in Healthcare. Computers in Biology and Medicine 2023, 164, 107295–107295. https://doi.org/10.1016/j.compbiomed.2023.107295.

139. Foster, D. Generative Deep Learning; “O’Reilly Media, Inc.,” 2019.

140. 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.

141. Chien, J.-T. Deep Neural Network. In Source Separation and Machine Learning; Elsevier: Amsterdam, The Netherlands, 2019; pp. 259–320.

142. 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.

143. 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.

144. EMA. Guideline on Reporting the Results of Population Pharmacokinetic Analyses; Committee for Medicinal Products for Human Use (CHMP): Amsterdam, The Netherlands, 2007

Appendix A

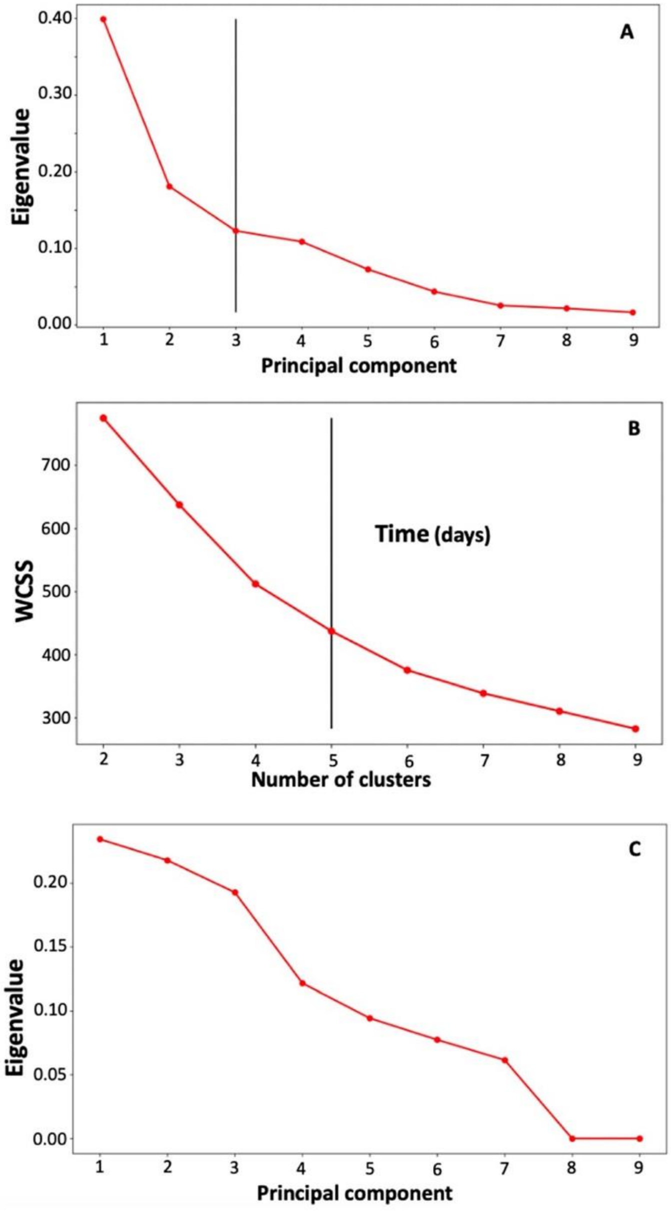


Figure A1. Scree (elbow) plots to determine the number of principal components or clusters in principal component analysis (A), k-means clustering (B), and factor analysis of mixed data (C). The x-axis refers to the number of principal components or clusters, while the y-axis corresponds to eigenvalues (for A,C) and the Within-Cluster Sum of Square (WCSS) (for plot B).

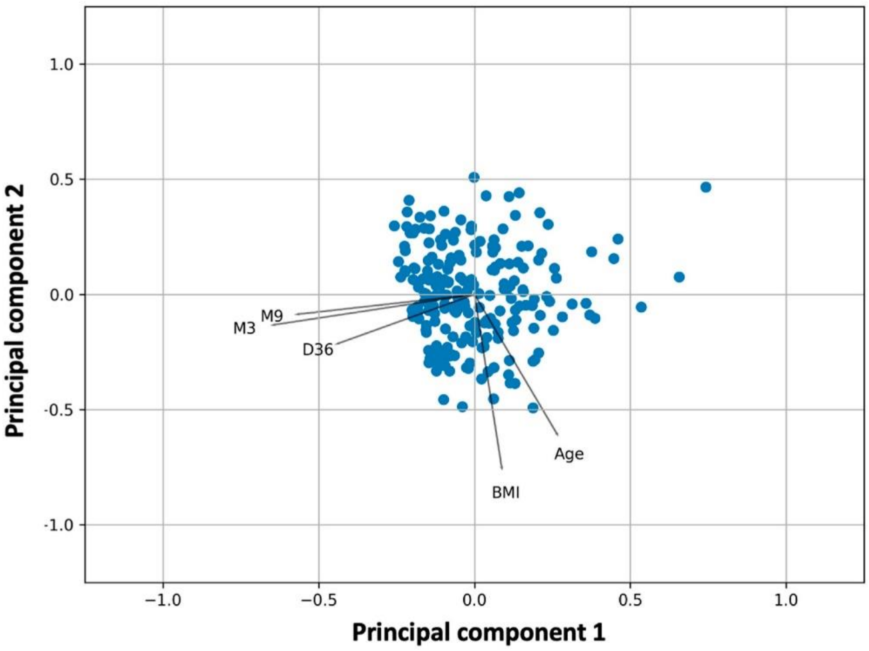


Figure A2. Principal component analysis of percent inhibition of SARS-CoV-2 binding at day 36 and at the third and ninth months after vaccination. The coordinates of the two-dimensional plot refer to the “scores” of the variables in the dimensionality reduced space (two principal components are shown). Key: D36, neutralizing antibody levels two weeks after second vaccination; M3 neutralizing antibody levels three months after second vaccination; M9, neutralizing antibody levels nine months after second vaccination; BMI, body mass index.

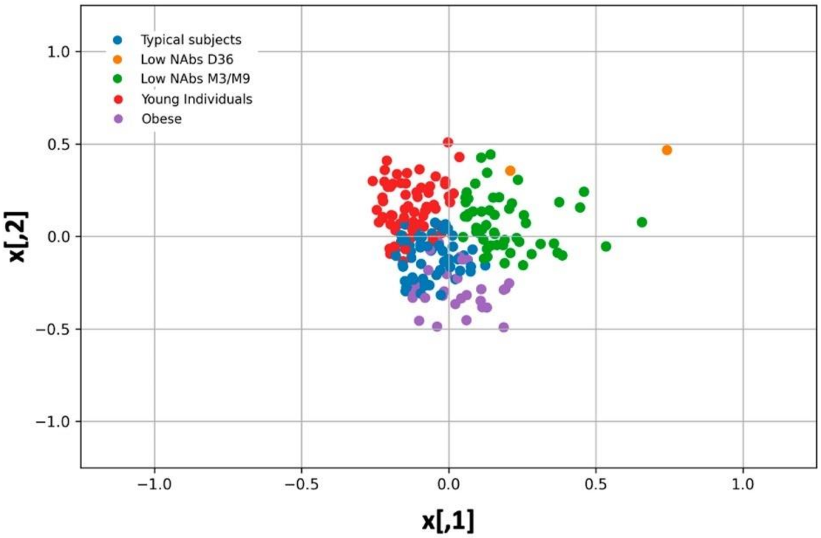


Figure A3. K-means cluster analysis plot of all features used in the study (neutralizing antibody levels and demographics). The two axes are labelled as x[,1] and x[,2]. In each cluster, shown with different color, a feature predominates.

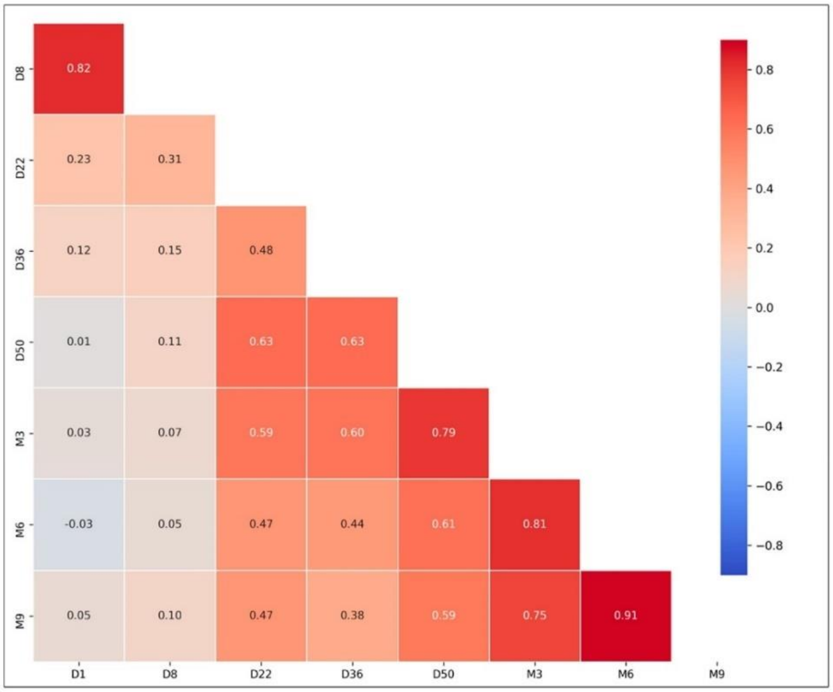
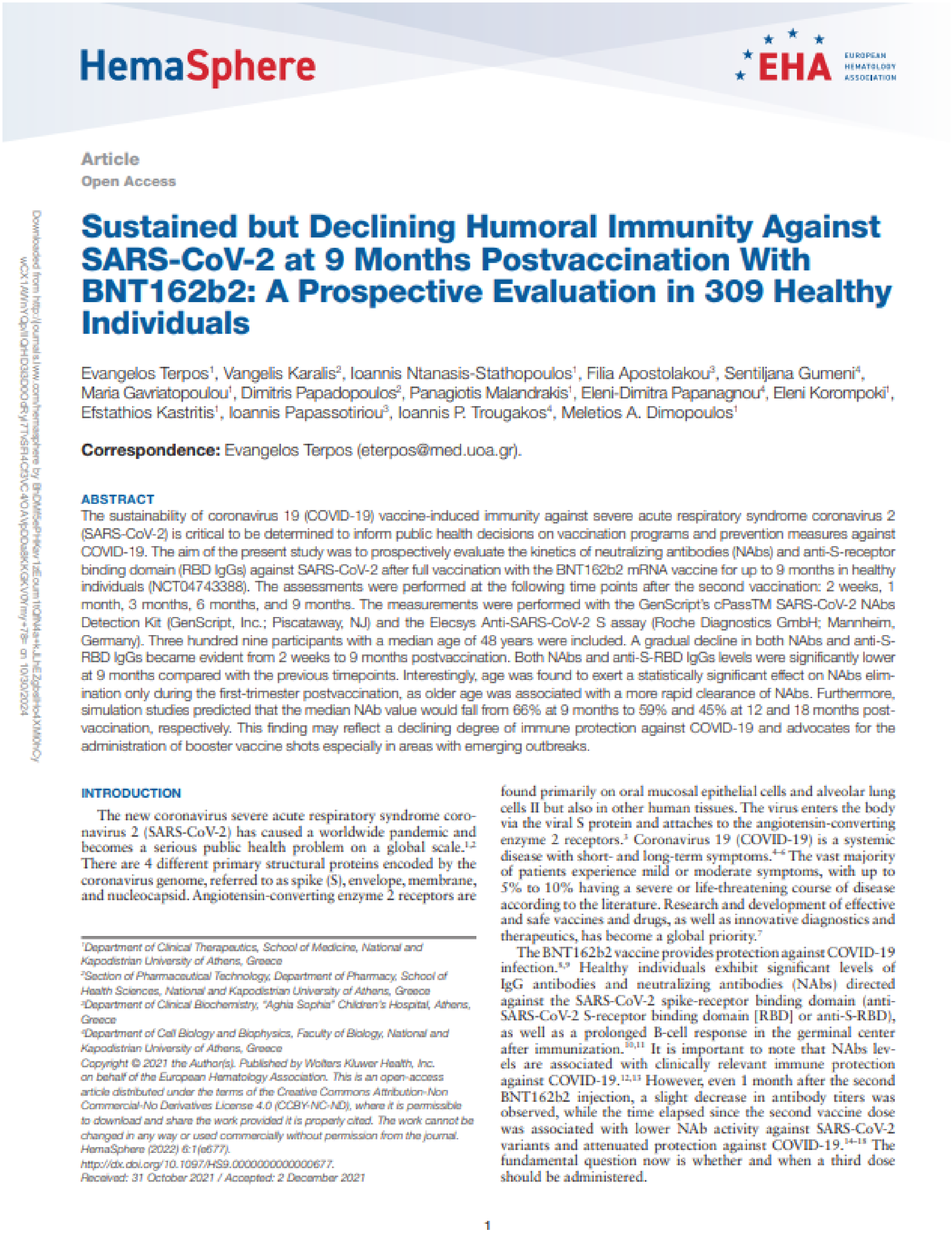


Figure A4. Correlation matrix showing the correlation coefficients (Spearman's) between the numerical variables used in the analysis.

Appendix B. Front covers of published papers



HemaSphere

Article
Open Access

Sustained but Declining Humoral Immunity Against SARS-CoV-2 at 9 Months Postvaccination With BNT162b2: A Prospective Evaluation in 309 Healthy Individuals

Evangelos Terpos[1], Vangelis Karalis[2], Ioannis Ntanasis-Stathopoulos[1], Filia Apostolakou[3], Sentiljana Gumeni[4], Maria Gavriatopoulou[1], Dimitris Papadopoulos[2], Panagiotis Malandrakis[1], Eleni-Dimitra Papanagnou[4], Eleni Korompoki[1], Efstathios Kastritis[1], Ioannis Papassotiriou[3], Ioannis P. Trougakos[4], Meletios A. Dimopoulos[1]

Correspondence: Evangelos Terpos (eterpos@med.uoa.gr).

ABSTRACT

The sustainability of coronavirus 19 (COVID-19) vaccine-induced immunity against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is critical to be determined to inform public health decisions on vaccination programs and prevention measures against COVID-19. The aim of the present study was to prospectively evaluate the kinetics of neutralizing antibodies (NAbs) and anti-S-receptor binding domain (RBD IgGs) against SARS-CoV-2 after full vaccination with the BNT162b2 mRNA vaccine for up to 9 months in healthy individuals (NCT04743388). The assessments were performed at the following time points after the second vaccination: 2 weeks, 1 month, 3 months, 6 months, and 9 months. The measurements were performed with the GenScript's cPassTM SARS-CoV-2 NAbs Detection Kit (GenScript, Inc.; Piscataway, NJ) and the Elecsys Anti-SARS-CoV-2 S assay (Roche Diagnostics GmbH; Mannheim, Germany). Three hundred nine participants with a median age of 48 years were included. A gradual decline in both NAbs and anti-S-RBD IgGs became evident from 2 weeks to 9 months postvaccination. Both NAbs and anti-S-RBD IgGs levels were significantly lower at 9 months compared with the previous timepoints. Interestingly, age was found to exert a statistically significant effect on NAbs elimination only during the first-trimester postvaccination, as older age was associated with a more rapid clearance of NAbs. Furthermore, simulation studies predicted that the median NAb value would fall from 66% at 9 months to 59% and 45% at 12 and 18 months postvaccination, respectively. This finding may reflect a declining degree of immune protection against COVID-19 and advocates for the administration of booster vaccine shots especially in areas with emerging outbreaks.

INTRODUCTION

The new coronavirus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused a worldwide pandemic and becomes a serious public health problem on a global scale.[1,2] There are 4 different primary structural proteins encoded by the coronavirus genome, referred to as spike (S), envelope, membrane, and nucleocapsid. Angiotensin-converting enzyme 2 receptors are found primarily on oral mucosal epithelial cells and alveolar lung cells II but also in other human tissues. The virus enters the body via the viral S protein and attaches to the angiotensin-converting enzyme 2 receptors.[3] Coronavirus 19 (COVID-19) is a systemic disease with short- and long-term symptoms.[4–6] The vast majority of patients experience mild or moderate symptoms, with up to 5% to 10% having a severe or life-threatening course of disease according to the literature. Research and development of effective and safe vaccines and drugs, as well as innovative diagnostics and therapeutics, has become a global priority.[7]

The BNT162b2 vaccine provides protection against COVID-19 infection.[8,9] Healthy individuals exhibit significant levels of IgG antibodies and neutralizing antibodies (NAbs) directed against the SARS-CoV-2 spike-receptor binding domain (anti-SARS-CoV-2 S-receptor binding domain [RBD] or anti-S-RBD), as well as a prolonged B-cell response in the germinal center after immunization.[10,11] It is important to note that NAbs levels are associated with clinically relevant immune protection against COVID-19.[12,13] However, even 1 month after the second BNT162b2 injection, a slight decrease in antibody titers was observed, while the time elapsed since the second vaccine dose was associated with lower NAb activity against SARS-CoV-2 variants and attenuated protection against COVID-19.[14–18] The fundamental question now is whether and when a third dose should be administered.

[1]Department of Clinical Therapeutics, School of Medicine, National and Kapodistrian University of Athens, Greece
[2]Section of Pharmaceutical Technology, Department of Pharmacy, School of Health Sciences, National and Kapodistrian University of Athens, Greece
[3]Department of Clinical Biochemistry, "Aghia Sophia" Children's Hospital, Athens, Greece
[4]Department of Cell Biology and Biophysics, Faculty of Biology, National and Kapodistrian University of Athens, Greece

HemaSphere (2022) 6:1(e677).
http://dx.doi.org/10.1097/HS9.0000000000000677.
Received: 31 October 2021 / Accepted: 2 December 2021

biomedicines

MDPI

Article

Predictive Factors for Neutralizing Antibody Levels Nine Months after Full Vaccination with BNT162b2: Results of a Machine Learning Analysis

Dimitris Papadopoulos [1], Ioannis Ntanasis-Stathopoulos [2], Maria Gavriatopoulou [2], Zoi Evangelakou [3], Panagiotis Malandrakis [2], Maria S. Manola [3], Despoina D. Gianniou [3], Efstathios Kastritis [2], Ioannis P. Trougakos [3], Meletios A. Dimopoulos [2], Vangelis Karalis [1,*,†] and Evangelos Terpos [2,*,†]

[1] Section of Pharmaceutical Technology, Department of Pharmacy, School of Health Sciences, National and Kapodistrian University of Athens, 15784 Athens, Greece; d.papadopoulos89@gmail.com

[2] Department of Clinical Therapeutics, School of Medicine, National and Kapodistrian University of Athens, 11528 Athens, Greece; johnntanasis@med.uoa.gr (I.N.-S.); mgavria@med.uoa.gr (M.G.); panosmalan@med.uoa.gr (P.M.); ekastritis@med.uoa.gr (E.K.); mdimop@med.uoa.gr (M.A.D.)

[3] Department of Cell Biology and Biophysics, Faculty of Biology, National and Kapodistrian University of Athens, 15784 Athens, Greece; zoievag@biol.uoa.gr (Z.E.); mmanola@biol.uoa.gr (M.S.M.); gndespoina@biol.uoa.gr (D.D.G.); itrougakos@biol.uoa.gr (I.P.T.)

* Correspondence: vkaralis@pharm.uoa.gr (V.K.); eterpos@med.uoa.gr (E.T.)

† These authors contributed equally to this work.

Abstract: Vaccination against SARS-CoV-2 with BNT162b2 mRNA vaccine plays a critical role in COVID-19 prevention. Although BNT162b2 is highly effective against COVID-19, a time-dependent decrease in neutralizing antibodies (NAbs) is observed. The aim of this study was to identify the individual features that may predict NAbs levels after vaccination. Machine learning techniques were applied to data from 302 subjects. Principal component analysis (PCA), factor analysis of mixed data (FAMD), k-means clustering, and random forest were used. PCA and FAMD showed that younger subjects had higher levels of neutralizing antibodies than older subjects. The effect of age is strongest near the vaccination date and appears to decrease with time. Obesity was associated with lower antibody response. Gender had no effect on NAbs at nine months, but there was a modest association at earlier time points. Participants with autoimmune disease had lower inhibitory levels than participants without autoimmune disease. K-Means clustering showed the natural grouping of subjects into five categories in which the characteristics of some individuals predominated. Random forest allowed the characteristics to be ordered by importance. Older age, higher body mass index, and the presence of autoimmune diseases had negative effects on the development of NAbs against SARS-CoV-2, nine months after full vaccination.

Keywords: SARS-CoV-2; COVID-19; neutralizing antibodies; machine learning; principal component analysis; factor analysis of mixed data; k-means clustering; random forest

check for updates

Citation: Papadopoulos, D.; Ntanasis-Stathopoulos, I.; Gavriatopoulou, M.; Evangelakou, Z.; Malandrakis, P.; Manola, M.S.; Gianniou, D.D.; Kastritis, E.; Trougakos, I.P.; Dimopoulos, M.A.; et al. Predictive Factors for Neutralizing Antibody Levels Nine Months after Full Vaccination with BNT162b2: Results of a Machine Learning Analysis. *Biomedicines* **2022**, *10*, 204. https://doi.org/10.3390/biomedicines10020204

Academic Editors: Yu-Te Wu and Wan-Yuo Guo

Received: 14 December 2021
Accepted: 17 January 2022
Published: 18 January 2022

1. Introduction

The novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has caused a worldwide epidemic and has become a serious global public health threat [1,2]. The coronavirus genome encodes four different major structural proteins: spike, envelope, membrane, and nucleocapsid. ACE2 receptors are typically found on epithelial cells of the oral mucosa and alveolar lung cells II but can also be found in other human organs [3]. The virus enters the body via the viral S protein and binds to ACE2 receptors [3]. COVID-19 is a systemic disease that causes both short- and long-term symptoms [4–6]. According to the literature, the vast majority of affected individuals have mild to moderate symptoms, with 5% to 10% having a severe or life-threatening course of disease. Worldwide, the

applied sciences

MDPI

Article

Variational Autoencoders for Data Augmentation in Clinical Studies

Dimitris Papadopoulos [1] 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.

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

check for updates

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

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

applied sciences

MDPI

Article

Introducing an Artificial Neural Network for Virtually Increasing the Sample Size of Bioequivalence Studies

Dimitris Papadopoulos [1] and Vangelis D. Karalis [1,2,*]

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

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

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

Abstract: Sample size is a key factor in bioequivalence and clinical trials. An appropriately large sample is necessary to gain valuable insights into a designated population. However, large sample sizes lead to increased human exposure, costs, and a longer time for completion. In a previous study, we introduced the idea of using variational autoencoders (VAEs), a type of artificial neural network, to synthetically create in clinical studies. In this work, we further elaborate on this idea and expand it in the field of bioequivalence (BE) studies. A computational methodology was developed, combining Monte Carlo simulations of 2 × 2 crossover BE trials with deep learning algorithms, specifically VAEs. Various scenarios, including variability levels, the actual sample size, the VAE-generated sample size, and the difference in performance between the two pharmaceutical products under comparison, were explored. All simulations showed that incorporating AI generative algorithms for creating virtual populations in BE trials has many advantages, as less actual human data can be used to achieve similar, and even better, results. Overall, this work shows how the application of generative AI algorithms, like VAEs, in clinical/bioequivalence studies can be a modern tool to significantly reduce human exposure, costs, and trial completion time.

Keywords: variational autoencoders; bioequivalence studies; artificial neural networks; sample size; clinical trials

check for updates

Citation: Papadopoulos, D.; Karalis, V.D. Introducing an Artificial Neural Network for Virtually Increasing the Sample Size of Bioequivalence Studies. *Appl. Sci.* **2024**, *14*, 2970. https://doi.org/10.3390/app14072970

Academic Editor: Aleksander Mendyk

Received: 8 February 2024
Revised: 30 March 2024
Accepted: 30 March 2024
Published: 31 March 2024

1. Introduction

Sample size estimation in clinical trials is a critical step which requires special attention, since any negligence in its calculation may result into misleading conclusions and can compromise the safety and the efficacy of the clinical trial [1,2]. A sufficiently large representative sample allows the derivation of robust insights for a given population. Although, collecting substantial amounts of data can be difficult, expensive, and long-lasting. Moreover, it is mandatory that a protocol is developed, which is followed by the clinical trial, in which the objectives, the primary and secondary endpoints, and all other aspects of the trial are clearly defined [1].

Sample size determination depends on the design of the study, the type of outcome, the statistical hypotheses, the expected variability in the measurement, the minimum detectable difference, the statistical power, and the level of significance [3]. Insufficient sample size in clinical trials could lead to an increase in type II error (false negatives), thus being unable to detect a true difference among groups and label the difference as statistically insignificant. In contrast, unnecessary sample size may be considered unethical or unfeasible due to high costs. Furthermore, all governmental and regulatory agencies worldwide require justification of the number of volunteers enrolled in the study.

As with any clinical trial, similar concerns apply to bioequivalence (BE) studies, especially in cases where a generic pharmaceutical product (i.e., Test, T) is compared against the reference product (R) [4,5]. Two pharmaceutical products are deemed bioequivalent if

applied sciences

MDPI

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.gr
[2] Department of Mathematics and Applied Mathematics, University of Crete, 71003 Heraklion, Greece
[3] 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: 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

check for updates

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

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

List of figures

List of tables

List of abbreviations

AEs	Autoencoders
AI	Artificial Intelligence
BE	Bioequivalence
CI	Confidence Interval
CNNs	Convolutional Neural Networks
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DL	Deep Learning
FAMD	Factor Analysis of Mixed Data
FN	False Negatives
FP	False Positives
GANs	Generative Adversarial Networks
GBM	Gradient Boosting Machine
GMM	Gaussian Mixture Models
GRUs	Gated Recurrent Units
kNN	k-Nearest Neighbors
LightGBM	Light Gradient Boosting
LLMs	Large Language Models
LSTMs	Long Short-term Memory Networks
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NAbs	Neutralizing Antibodies
NLP	Natural Language Processing
NNs	Neural Networks
PC	Principal Component
PCA	Principal Components Analysis
RNNs	Recurrent Neural Networks
RSS	Residual Sum of Squares
SVD	Singular Value Decomposition
SVM	Support Vector Machines

SVR	Support Vector Regression
TN	True Negatives
TP	True Positives
t-SNE	t-distributed Stochastic Neighbor Embedding
VAEs	Variational autoencoders
WCSS	Within-cluster Sum of Squares
Xgboost	Extreme Gradient Boosting