pharmaceutics

Article

Alternative Analysis Approaches for the Assessment of Pilot Bioavailability/Bioequivalence Studies

Sara Carolina Henriques [1,2,*], João Albuquerque [2,3], Paulo Paixão [1], Luís Almeida [2] and Nuno Elvas Silva [1,*]

[1] Research Institute for Medicines (iMed.ULisboa), Faculty of Pharmacy, Universidade de Lisboa, 1649-003 Lisboa, Portugal; ppaixao@ff.ulisboa.pt
[2] BlueClinical Ltd., Senhora da Hora, 4460-439 Matosinhos, Portugal; joao.albuquerque@blueclinical.pt (J.A.); lalmeida@blueclinical.pt (L.A.)
[3] Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
* Correspondence: saracarolinahenriques@gmail.com (S.C.H.); nmens@ff.ulisboa.pt (N.E.S.)

Abstract: Pilot bioavailability/bioequivalence (BA/BE) studies are usually conducted and analysed similarly to pivotal studies. Their analysis and interpretation of results usually rely on the application of the average bioequivalence approach. However, due to the small study size, pilot studies are inarguably more sensitive to variability. The aim of this work is to propose alternative approaches to the average bioequivalence methodology, in a way to overcome and reduce the uncertainty on the conclusions of these studies and on the potential of test formulations. Several scenarios of pilot BA/BE crossover studies were simulated through population pharmacokinetic modelling. Each simulated BA/BE trial was analysed using the average bioequivalence approach. As alternative analyses, the centrality of the test-to-reference geometric least square means ratio (GMR), bootstrap bioequivalence analysis, and arithmetic (A_{mean}) and geometric (G_{mean}) mean f_2 factor approaches were investigated. Methods performance was measured with a confusion matrix. The G_{mean} f_2 factor using a cut-off of 35 was the most appropriate method in the simulation conditions frame, enabling to more accurately conclude the potential of test formulations, with a reduced sample size. For simplification, a decision tree is also proposed for appropriate planning of the sample size and subsequent analysis approach to be followed in pilot BA/BE trials.

Keywords: bioequivalence; generic medicinal products; pilot studies; f_2 factor; bootstrap; pharmacokinetics; pharmacokinetic simulation

check for updates

Citation: Henriques, S.C.; Albuquerque, J.; Paixão, P.; Almeida, L.; Silva, N.E. Alternative Analysis Approaches for the Assessment of Pilot Bioavailability/Bioequivalence Studies. *Pharmaceutics* **2023**, *15*, 1430. https://doi.org/10.3390/pharmaceutics15051430

Academic Editor: Alain Hedoux

Received: 7 April 2023
Revised: 3 May 2023
Accepted: 5 May 2023
Published: 7 May 2023

1. Introduction

Bioavailability and bioequivalence are the cornerstone for the approval of brand-name and generic drugs globally under the European Medicines Agency (EMA) [1] and US Food and Drug Administration (FDA) [2] ambience. According to EMA, bioequivalence is the absence of a significant difference in the bioavailability (i.e., rate and extent) to which the active substance in pharmaceutical equivalents or pharmaceutical alternatives becomes available at the site of drug action when administered at the same molar dose, under similar conditions [1]. When bioequivalence between two drug products is claimed, an equivalent therapeutic efficacy and safety are assumed. Therefore, lengthy and costly phase III clinical trials on the bioequivalent test product may be waived [3].

A key goal in pharmaceutical development of oral dosage forms is a good understanding of the in vivo and in vitro performance of the dosage form and the optimization of an in vitro profile for the potential formulation that reflects its in vivo performance. In vitro dissolution testing provides useful information at several stages of the drug product development process and is usually used to assist scientists on excipients selection and manufacturing process adjustments that originate a candidate formulation with the most suitable and reproducible release profile. Therefore, dissolution results are commonly the decision key to test the new formulation in vivo. Nevertheless, dissolution results not always guarantee a correlation between the in vitro and the in vivo performance of the developed formulations [4]. Knowing this, and following a conservative approach, it is usual for companies to carry out pilot bioavailability/bioequivalence (BA/BE) studies.

A pilot BA/BE study is a downsized trial that can be conducted prior to the definitive pivotal trial and may act as a gatekeeping in vivo strategy to decide whether to move forward with a full-size pivotal study [4,5]. The pilot study can serve as a valuable tool (i) to validate the analytical methodology, (ii) to assess pharmacokinetic variability and to determine sample size to achieve adequate power, (iii) to optimize sample collection time intervals, (iv) to determine the needed washout period between treatments, and (v) to gather information about the formulation (or formulations) being tested against a reference product, and to assess its eligibility as a possible bioequivalent candidate(s) [1,2,5,6].

Pilot studies are usually conducted and analysed similarly to pivotal studies. Literature and guidelines provide no formal methodologies, besides the application of an average bioequivalence [4,6]. Average bioequivalence is a parametric approach based solely on the comparison of means, whether other characteristics of the distributions of the selected bioavailability metrics (e.g., inter- or intra-subject variabilities) are ignored. For formal pivotal BA/BE studies, test and reference formulations can be considered bioequivalent if the 90% confidence interval (CI) of the ratio of the geometric least square means of the pharmacokinetic parameters of interest are within the acceptance interval of [80.00–125.00]% [1,2,7].

The number of subjects to be included in a pilot study is generally 12–18 subjects, depending on the expected intra-subject coefficient of variation (ISCV%) [1,2]. However, due to their small sample size, pilot studies are inarguably more sensitive to variability. The point estimate obtained for the means ratio may not be close to the real population value, particularly when variability is high. Consequently, there is a greater risk of either (i) validating a bioinequivalent test formulation and proceeding further with a pivotal study or (ii) discarding a potentially bioequivalent formulation, by not conducting a pivotal study. Moreover, due to the small sample size, the 90% CI of the mean may be much wider and therefore fall out of the acceptance interval and reduce the probability of a positive decision.

The aim of this work is to propose alternative approaches to the average bioequivalence methodology that is generally applied to pilot studies, in a way to reduce the uncertainty on the conclusions of these studies and on the potential of test formulations. Several scenarios of pilot BA/BE crossover studies were simulated through population pharmacokinetic modelling, accounting for different inter-individual (IIV) and inter-occasion (IOV) levels of variability. Methods performance was measured with a confusion matrix.

2. Materials and Methods

A total of 32,000 BA/BE crossover trials (corresponding to 1,344,000 different concentration-time profiles) were simulated (i) accounting for different sample sizes, (ii) combining different IIV and/or IOV variability levels for the pharmacokinetic parameters, (iii) and considering no difference or a difference between test and reference products on the mean absorption rate constant (k_a) (Figure 1).

Trial simulations and statistical analysis were performed with R version 4.0.3 (R Foundation for Scientific Computing, 2013).

2.1. Study Design

All studies were simulated as two-sequence (Sequence 1 and Sequence 2), two-treatment (test and reference), two-period crossover (2 × 2 × 2) studies, accounting for a range of 12–30 (in an increment of two) subjects. Subjects were randomized prior to pharmacokinetic simulation. A computer-generated balanced block-wise randomization list was appropriately generated according to the study sample size (Figure 1—Study Design).

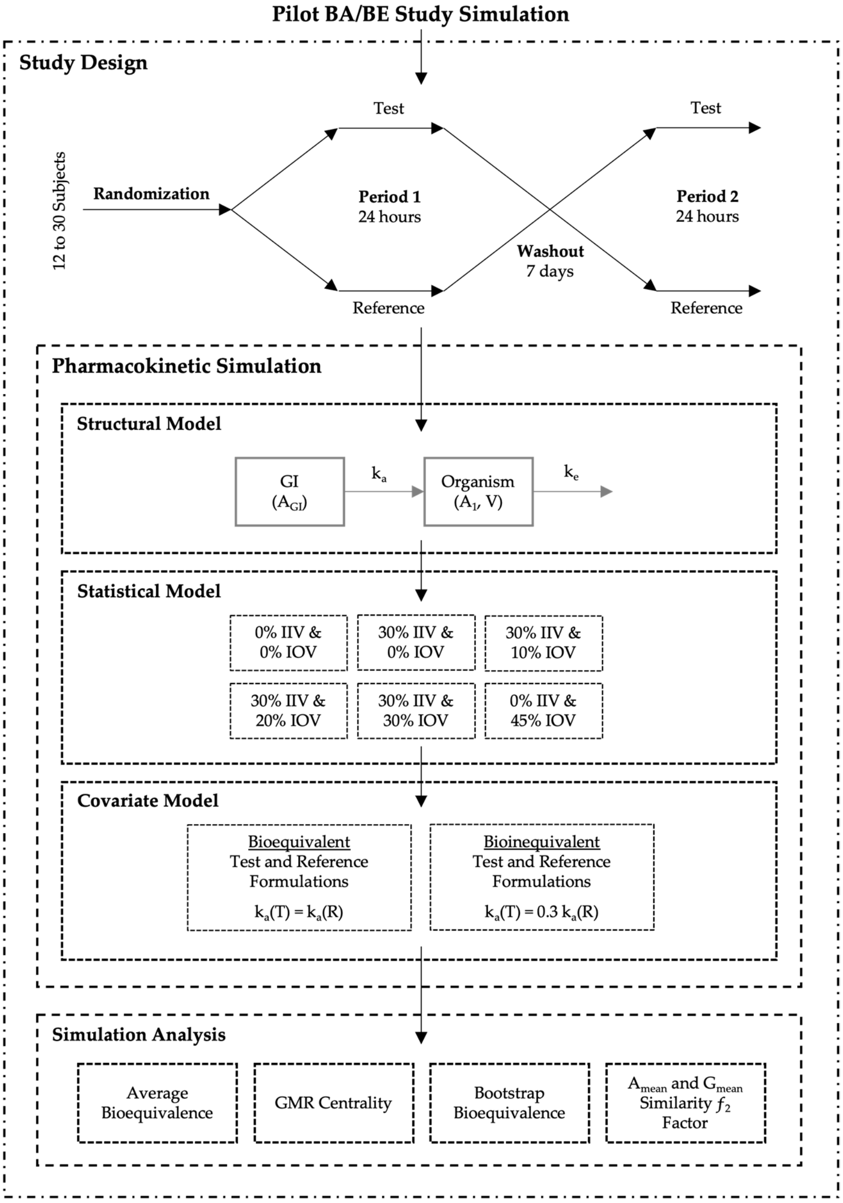

Figure 1. Pilot bioavailability/bioequivalence (BA/BE) trials simulation scheme.

2.2. Population Pharmacokinetic Simulation

For the simulation of plasma concentration-time profiles, a population pharmacokinetic modelling and simulation approach was used.

A one-compartmental model with first-order absorption and first-order elimination was selected as the simplest oral model to describe the processes of drug absorption and disposition. Simulations were performed through ordinary differential equations (ODE), parameterized with micro constants (Figure 1—Structural Model), and Equations (1) and (2), where A_{GI} represents the amount in the gastrointestinal tract, A_1 the amount in the organism, k_a the absorption rate constant, k_e the elimination rate constant, V the apparent volume of distribution and C the plasma concentration.

$$\begin{cases} \frac{dA_{GI}}{dt} = -k_a \cdot A_{GI} \\ \frac{dA_1}{dt} = k_a \cdot A_{GI} - k_e \cdot A_1 \end{cases} \tag{1}$$

$$C = \frac{A_1}{V} \tag{2}$$

A range of 12–30 subjects per study was simulated using 'Simulx', a function of the 'mlxR' package version 4.1.3 (Monolix version 2019R2, Lixoft, Antony, France), that implements ODEs based mixed effects models by interfacing the C++ MlxLibrary with R. According to the previously defined sequence-balanced randomization scheme, subjects were administered a single 50 mg oral dose of either test or reference products, separated by a washout of 7 days (Figure 1—Study Design). Each pharmacokinetic profile comprised 20 simulated plasma samples, at the time of dose (time 0); and at 0.25, 0.50, 0.75, 1.00, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00, 3.25, 3.50, 3.75, 4.00, 6.00, 8.00, 12.00, and 24.00 h after dose. For each individual and occasion, compartmental pharmacokinetic parameters were generated considering one of the following six (6) different variability scenarios: (i) baseline (0% IIV and 0% IOV), (ii) 30% IIV and 0% IOV, (iii) 30% IIV and 10% IOV, (iv) 30% IIV and 20% IOV, (v) 30% IIV and 30% IOV, and (vi) 0% IIV and 45% IOV (Figure 1—Statistical Model). Each variability scenario was applied separately to each model parameter, i.e., variability scenarios were not applied simultaneously. Mean values for k_a, V, and k_e are presented in Table 1. All parameters followed a log-normal distribution (Equation (3)). Absolute bioavailability (F) was considered to have a mean value of 0.9 (Table 1), and no variability was tested for this parameter.

$$\begin{gathered}\log(\Psi_i) \sim \mathcal{N}\left(\log\left(\overline{\Psi}_i\right), \left(\omega^2, \gamma^2\right)\right) \\ \Leftrightarrow \\ \Psi_i = \overline{\Psi}_i \cdot e^{\eta_i+\kappa_i}, \text{ where } \eta_i \sim \mathcal{N}\left(0, \omega^2\right) \text{ and } \kappa_i \sim \mathcal{N}\left(0, \gamma^2\right)\end{gathered} \tag{3}$$

Table 1. Compartmental Pharmacokinetic Parameters Initial Estimates.

k_a **(h^{-1})**	**V** **(L)**	**k_e** **(h^{-1})**	**F**
1.22	58.8	0.150	0.900

F: Absolute bioavailability, k_a: absorption rate constant, k_e: elimination rate constant, V: volume of distribution.

Furthermore, the individual plasma concentrations over time were simulated considering a log-normal additive experimental error (Equation (4)). A coefficient of variation (CV%) of 10% was used to reflect sampling and quantification errors. For simplicity, neither sequence nor period variability was included.

$$\begin{gathered}Y = f(\theta; x) \cdot e^{\varepsilon} \\ \Leftrightarrow \\ \log Y = \log f(\theta; x) + \varepsilon\end{gathered} \tag{4}$$

Moreover, for the purpose of this work, two groups of simulations were planned: one group where test and reference products were considered equal, with the same mean values for the pharmacokinetic model parameters (i.e., truly bioequivalent), and another group where test product presented a mean k_a value as 30% of the reference product mean k_a (i.e., truly bioinequivalent) (Figure 1—Covariate Model).

Within each group of simulations, and for each variability scenario, 100 bioequivalence crossover trials were simulated. The pharmacokinetic parameters maximum observed plasma concentration (C_{max}) and area under the plasma concentration-time curve (AUC) were derived for each pharmacokinetic profile. The AUC typically reflects the extent of drug absorption, whether C_{max} is considered to reflect the absorption rate [1,2]. C_{max} usually shows larger variation compared to AUC, as the parameter highly depends on the selection of sampling times. Thus, as the risk of failing to demonstrate bioequivalence is higher for the rate of drug absorption, performed simulations only covered the effect of variability on the bioequivalence of C_{max}.

2.3. Simulations Analysis

Each simulated bioequivalence trial was analysed using the average bioequivalence approach. As alternatives, the centrality of the test-to-reference GMR, a bootstrap bioequivalence analysis, and arithmetic (A_{mean}) and geometric (G_{mean}) mean f_2 factor approaches were also investigated.

2.3.1. Average Bioequivalence Analysis

An analysis of variance (ANOVA) was performed on the *ln*-transformed C_{max}. A linear model was applied, using sequence, subject nested within sequence, period and treatment as fixed effects [1,8,9].

As in accordance with EMA's Guideline on the Investigation of Bioequivalence [1] and US FDA Guidance [2,7], for each simulated pilot study, the assessment of bioequivalence was based upon the 90% CI for the test-to-reference geometric least square means (LSM) ratio (GMR) for the primary pharmacokinetic parameter. This method is based on Schuirmann's two one-sided *t*-tests (TOST) with the null hypothesis of bioinequivalence at the 5% significance level (α = 0.05) [1–3,7,10]. Assuming a maximum 20% difference between test and reference formulations, the interval hypotheses for average bioequivalence can be formulated as

$$\begin{gathered} \mathrm{H_0}:\ \mu_\mathrm{T}-\mu_\mathrm{R} < ln(0.80) \text{ or } \mu_\mathrm{T}-\mu_\mathrm{R} > -ln(0.80) \\ \text{versus } \mathrm{H_1}:\ ln(0.80) \le \mu_\mathrm{T}-\mu_\mathrm{R} \le -ln(0.80) \end{gathered} \tag{5}$$

where μ_T and μ_R are the population average response (i.e., the LSM) of the *ln*-transformed measure for test and reference formulations, respectively. Hence, for the back-transformed data, the hypotheses for average bioequivalence can be expressed as

$$\begin{gathered} \mathrm{H_0}:\ e^{\mu_\mathrm{T}}/e^{\mu_\mathrm{R}} < 0.80 \text{ or } e^{\mu_\mathrm{T}}/e^{\mu_\mathrm{R}} > 1.25 \\ \text{versus } \mathrm{H_1}:\ 0.80 \le e^{\mu_\mathrm{T}}/e^{\mu_\mathrm{R}} \le 1.25 \end{gathered} \tag{6}$$

where the alternative hypothesis (H_1) is shown by rejecting the null hypothesis (H_0) of average bioinequivalence, i.e., the decision of bioequivalence is based on whether the 90% CI ($100\cdot(1-2\alpha)$%) of the test-to-reference GMR is within the regulatory acceptance interval of [80.00–125.00]% [3,7,10,11].

Moreover, the ISCV% was estimated for each of the primary pharmacokinetic parameters, as

$$\mathrm{ISCV\%} = 100 \cdot \sqrt{e^{s^2}-1} \tag{7}$$

where s^2 is the mean square error obtained from the ANOVA model of the *ln*-transformed parameters [1,8,9].

Average bioequivalence analysis was performed through in-house functions developed in R, previously validated with Phoenix® WinNonlin® version 8.3 (Certara USA Inc., Princeton, NJ, USA).

2.3.2. Centrality of the Test-to-Reference GMR

Beyond the standard average bioequivalence approach, the centrality of the test-to-reference GMR was tested, i.e., for each simulated pilot study it was verified if the attained test-to-reference GMR was within the tighter acceptance range of [90.00–111.11]%.

2.3.3. Bootstrap Bioequivalence Analysis

Bootstrapping is a non-parametric method that can be used to assess the precision of a statistic without making strong assumption for the distribution from which samples are drawn [12].

Using Monte Carlo simulations, new sets of pharmacokinetic data were created, by repeatedly sampling from the simulated study data with replacement. By resampling with replacement, the bootstrap resampling mimics the experimental procedure [12].

For each simulated study, a total of 100 bootstrap resamples were generated from the simulated concentration-time profiles [12,13]. The sample size of the bootstrap resamples was calculated based on the original simulated study, from which the average bioequivalence approach was applied and ISCV% was estimated. For sample size calculation for bootstrap resampling, it was assumed the estimated ISCV%, a power of 80%, a true test-to-reference GMR of 90% and an α of 0.05, using R package 'PowerTOST' version 1.5–3 [14]. To ensure that the number of resampled subjects assigned with sequences 1 and 2 was the same, resampling was sequence balanced.

From the resampled data set, average bioequivalence was re-computed, thus generating a bootstrap estimate of the statistics of interest. The bootstrap resamples' GMR were then used to estimate the standard error of the bootstrap GMR and its corresponding 95% CI. The non-parametric confidence bounds were obtained as percentiles from the bootstrap estimator of the sampling distribution of the parameter estimator [12,15]. A was reduced to 0.025, in order to better circumvent type I errors. Similarly, the decision of bioequivalence was based on whether the bootstrap 95% CI of the test-to-reference GMR was within the regulatory acceptance interval of [80.00–125.00]%.

2.3.4. Similarity f_2 Factor

The similarity f_2 factor is a mathematical index widely used to compare dissolution profiles, evaluating their similarity, using the percentage of drug dissolved per unit of time. The similarity f_2 factor, proposed by Moore and Flanner in 1996 [16], is derived from the mean squared difference, and can be calculated as a function of the reciprocal of mean squared-root transformation of the sum of square differences at all points:

$$f_2 = 50 \cdot \log\left(100 \cdot \left[1 + \frac{1}{n}\sum_{t=1}^{t=n}\left(\overline{\mathrm{R}}_t - \overline{\mathrm{T}}_t\right)^2\right]^{-0.5}\right) \tag{8}$$

where f_2 is the similarity factor, n is the number of time points, and $\overline{\mathrm{R}}_t$ and $\overline{\mathrm{T}}_t$ are the mean percentage of drug dissolved at time t after initiation of the study, for reference and test products, respectively [1,16,17].

The f_2 similarity factor ranges from 0 (when $\overline{\mathrm{R}}_t - \overline{\mathrm{T}}_t = 100\%$, at all t) to 100 (when $\overline{\mathrm{R}}_t - \overline{\mathrm{T}}_t = 0\%$, at all t). Therefore, applying Equation (8), an average difference of 10%, 15%, and 20% from all measured time points results in a f_2 value of 50, 41, and 35, respectively (Figure 2). EMA [1] and FDA [18,19] have set a public standard of f_2 value between 50–100, i.e., a maximum mean difference of 10%, to indicate similarity between the two dissolution profiles.

In this work, the concept of similarity factor f_2 was applied as an alternative to the average bioequivalence analysis. The similarity between test and reference products by means of f_2 was evaluated through the comparison of arithmetic ($\mathrm{A_{mean}}$) and geometric ($\mathrm{G_{mean}}$) means of plasma concentration-time profiles derived from the simulated individual pharmacokinetic profiles. f_2 was used to assess the similarity on the rate of drug absorption by normalizing test and reference mean concentration-time profiles to the maximum plasma concentration ($\mathrm{C_{max}}$) derived from the mean reference profile, until reference $\mathrm{C_{max}}$ is observed (reference $\mathrm{t_{max}}$) (Equation (9)).

$$\mathrm{C}_t^N = 100 \cdot \frac{\overline{\mathrm{C}}_t}{\mathrm{C}_{\max R}}, \quad \text{where } 0 \leq \mathrm{t} \leq \mathrm{t}_{\max R} \tag{9}$$

In Equation (9), C_t^N is the normalized concentration at time t, $\overline{C}_t$ is the mean (test or reference) concentration at time t, $C_{\max R}$ is the $C_{\max}$ of the reference mean concentration-time profile, and $t_{\max R}$ the time of observation of $C_{\max R}$. The similarity f_2 factor is calculated as

$$C_{\max}\ f_2 = 50\cdot \log\left(100\cdot\left[1+\frac{1}{n}\sum_{t=1}^{t=n}\left(R_t^N - T_t^N\right)^2\right]^{-0.5}\right) \tag{10}$$

where n is the number of time points until reference $t_{\max}$, and R_t^N and T_t^N are the normalized concentration at time t for reference and test products, respectively.

Then, the f_2 factor was tested for differences between test and reference formulations' mean concentration-time profiles of 10%, 15% and 20%. Consequently, the interval hypotheses for the f_2 factor can be formulated as

$$\begin{gathered} H_0:\ f_2 < \theta \\ \text{versus } H_1:\ f_2 \geq \theta. \end{gathered} \tag{11}$$

where θ is the tested cut-off of (i) 35 for testing maximum differences of 20%; (ii) 41 for differences of 15%; and (iii) 50 for testing differences of 10% between the concentration-time profiles.

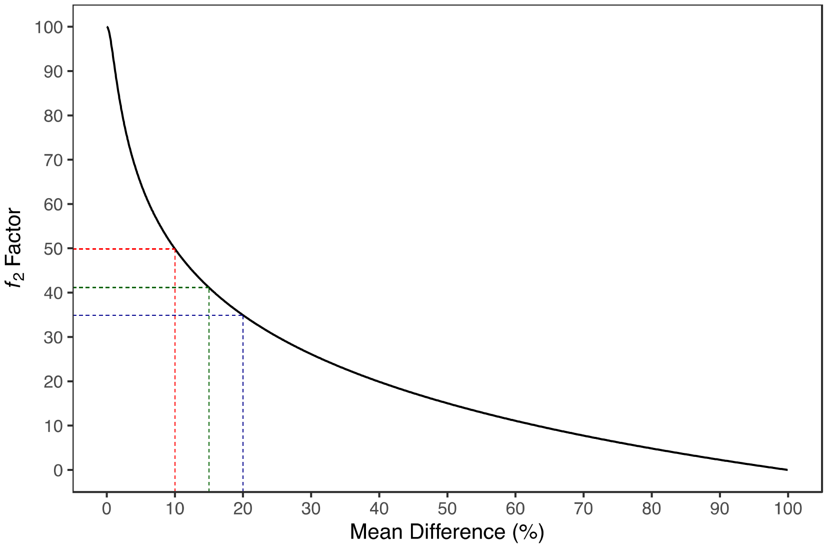

Figure 2. Distribution of f_2 similarity factor as a function of mean difference. f_2 similarity factor is derived from the mean squared difference and can be calculated as a function of the reciprocal of the mean squared-root transformation of the sum of square differences at all points. An average difference of 10%, 15%, and 20% from all measured time points results in a f_2 value of 50 (red dotted lines), 41 (green dotted lines) and 35 (blue dotted lines), respectively.

2.4. Performance Measurement

When testing a hypothesis, two errors may occur: (i) the type I error, which concerns the rejection of true H_0 (Equation (12)); (ii) and the type II error which concerns the failing to reject false H_0 (Equation (13)) [3]. The probabilities of making type I and type II errors are given as

$$\alpha = P(\text{type I error}) = P\left(\text{reject } H_0 \mid H_0 \text{ is true}\right) \tag{12}$$

$$\beta = P(\text{type II error}) = P\left(\text{fail to reject } H_0 \mid H_0 \text{ is false}\right) \tag{13}$$

In order to find the relationship between type I and type II errors, hence determining the performance of each bioequivalence evaluation method (average bioequivalence, centrality of the test-to-reference GMR, bootstrap bioequivalence, and A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35, 41, and 50), a confusion matrix, i.e., a cross-tabulation of the observed and predicted classes with associated statistics, was created (Table 2).

Table 2. Confusion Matrix of the Observed and Predicted Classes with Associated Statistics.

		Method Prediction		
		Bioequivalent	**Bioinequivalent**	
Truly	**Bioequivalent**	TP	FN Type II Error	Sensitivity $\frac{TP}{TP\ \ +FN}$
	Bioinequivalent	FP Type I Error	TN	Specificity $\frac{TN}{TN+FP}$
		Precision $\frac{TP}{TP+FP}$	Negative Predictive Value $\frac{TN}{TN+FN}$	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

FN—False Negatives; FP—False Positives, TN—True Negatives, TP—True Positives.

For each evaluation method, the created matrixes accommodated (i) the true positives (TP), i.e., the number of correctly identified bioequivalent predictions; (ii) the false negatives (FN), i.e., the number of incorrectly identified bioinequivalent predictions; (iii) the false positives (FP), i.e., the number of incorrectly identified bioequivalent predictions; and (iv) the true negatives (TN), i.e., the number of correctly identified bioinequivalent predictions (Table 2).

Moreover, the following statistics were derived from the cross-tabulated matrix [20]:

1. Sensitivity, also referred to as power, recall or true positive rate, which measures the capacity of the model to correctly identify bioequivalent test and reference formulations. In other words, it is the probability of correctly rejecting H_0 when H_0 is false (Table 2).

$$\text{Sensitivity } \textit{or} \text{ Power} = 1 - \beta = P\left(\text{reject } H_0 \;\middle|\; H_0 \text{ is false}\right) \tag{14}$$

When the test recognizes all the bioequivalent formulations (i.e., no false negatives) Sensitivity = 1; when the test does not recognize any of the bioequivalent formulations Sensitivity = 0.

2. Specificity, also referred to as true negative rate, measures the capacity of the model to correctly identify bioinequivalent test and reference formulations. In other words, it is the probability of correctly failing to reject H_0 when H_0 is true (Table 2).

$$\text{Specificity} = 1 - \alpha = P\left(\text{fail to reject } H_0 \;\middle|\; H_0 \text{ is true}\right) \tag{15}$$

When the test recognizes all the bioinequivalent formulations (i.e., no false positives) Specificity = 1; when the test does not recognize any of the bioinequivalent formulations Specificity = 0.

3. Precision, also referred to as positive predictive value (PPV), measures the correctness achieved in bioequivalent predictions (Table 2). When PPV = 1, all identified bioequivalent formulations are truly bioequivalent.
4. Negative Predictive Value (NPV), which measures the correctness achieved in bioinequivalent predictions (Table 2). When NPV = 1, all identified bioinequivalent formulations are truly bioinequivalent.
5. Accuracy, which represents the ratio between the correctly identified predicted instances (bioequivalent and bioinequivalent) and the total number of instances (Table 2). When Accuracy = 1, the test predicted correctly all the bioequivalent and bioinequivalent formulations.

6. F_1 score, which is the harmonic mean of Sensitivity and Precision.

$$F_1 = \frac{2 \cdot \text{Sensitivity} \cdot \text{Precision}}{\text{Sensitivity} + \text{Precision}} \tag{16}$$

   F_1 score is independent from the number of samples correctly classified as negative. A F_1 = 1 indicates perfect precision and sensitivity; for a F_1 = 0, either precision or sensitivity are 0.

7. Matthews Correlation Coefficient (MCC), which measures the correlation coefficient between the true classes and the method predicted classes.

$$\text{MCC} = \frac{Cov(t,p)}{\sigma_t \cdot \sigma_p} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP}+\text{FP}) \cdot (\text{TP}+\text{FN}) \cdot (\text{TN}+\text{FP}) \cdot (\text{TN}+\text{FN})}} \tag{17}$$

   where $Cov(t, p)$ is the covariance of the true classes t and predicted labels p, whereas σ_t and σ_p are the standard deviations, respectively [21]. A MCC = 1 indicates a perfect prediction; MCC = 0 indicates that the prediction is no better than random; and MCC = -1 indicates total disagreement between prediction and observation.

8. Cohen's Kappa (κ) statistic, which is a measure of concordance for categorical data that measures agreement relative to what would be expected by chance.

$$\kappa = \frac{2 \cdot (\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN})}{(\text{TP}+\text{FP}) \cdot (\text{TP}+\text{FN}) \cdot (\text{TN}+\text{FP}) \cdot (\text{TN}+\text{FN})} \tag{18}$$

   When there is complete agreement κ = 1; when there is no agreement κ = 0; and when there is no effective agreement, or when there is a complete disagreement, κ = -1.

3. Results

3.1. Simulated Pharmacokinetic Data

Histograms of the individual estimates of pharmacokinetic parameters exhibited centred distribution around the population's typical value. Descriptive statistics of all simulated pharmacokinetic parameters and a graphical representation of their distribution are presented in Appendix SA.

The defined sampling times were appropriate to describe the simulated concentration-time profiles (Appendix SA).

For the group of simulations concerning the test product as truly bioequivalent to the reference product, it was observed for the baseline simulations (0% IIV and 0% IOV) that the C_{max} for both test and reference products was approximately 642.5 μg/L (geometric coefficient of variation [GCV%] ≈ 6%), being reached between 0.75 and 4 h (median t_{max} = 2.25 h). AUC from pre-dose until the last sampling time was approximately 4950 μg.h/L (GCV% ≈ 3%). As expected, no differences between test and reference products were observed for these NCA parameters (Appendix SA.1.2).

For the group of simulations concerning the test product as truly bioinequivalent to the reference product, it was observed a delayed t_{max} for the test product (t_{max} = 3.5 h [1.75–8 h]), as well as a 30% reduction of C_{max} (G_{mean} = 460.69 μg/L [GCV% ≈ 6%]), as a consequence of the differences in k_a between test and reference products. No differences were observed for AUC, as test and reference products only presented differences in k_a, and not in F (Appendix SA.1.2).

The increment in the variability of k_a (IIV and IOV) did not greatly affect the distribution of C_{max} nor AUC values (GCV% ≈ 7–20%); however, it increased the time range for the observation of C_{max}. t_{max} values ranged from 0.5 to 6 h for the reference product and ranged from 0.25 to 8 h for a truly bioequivalent test, and from 1 to 12 h for a truly bioinequivalent test (Appendix SA.2.3).

Likewise, an increase of variability in k_e was not associated with higher dispersion of C_{max} values, but it was associated with a wider time range for t_{max} (from 0.5 to 8 h for reference and truly bioequivalent test; and from 1 to 12 h for truly bioinequivalent test). An increase variability in k_e induced an increased variability in AUC (GCV% ≈ 27–40%) (Appendix SA.4.3).

Moreover, an increase of variability in V was associated with a wider dispersion of C_{max} and AUC values (GCV% ≈ 30–48%). However, no differences were observed for t_{max} range (Appendix SA.3.3).

3.2. Bioequivalence Evaluation

Confusion matrix results were graphically presented over the number of subjects for sensitivity in Figures 3 and 4, for specificity in Figure 5, for precision in Figure 6, for NPV in Figure 7, for accuracy in Figure 8, for F_1 in Figure 9, for MCC in Figure 10 and for Cohen's κ in Figure 11.

Considering the evaluation of bioequivalence, as expected, the presence of type I and type II errors depended only on IOV, and not on IIV, as IIV is suppressed by using a crossover design.

For the pharmacokinetic model tested, IOV in V was the variability identified with the highest impact on the evaluation of C_{max} bioequivalence metric. The variability tested for the other model parameters had no relevant impact.

Variability (IOV) in V turns critical for bioequivalence evaluation at 20% IOV, a level of variability lower than the established cut-off value of 30% that defines a highly variable drug/drug product, in accordance with EMA's Guideline [1].

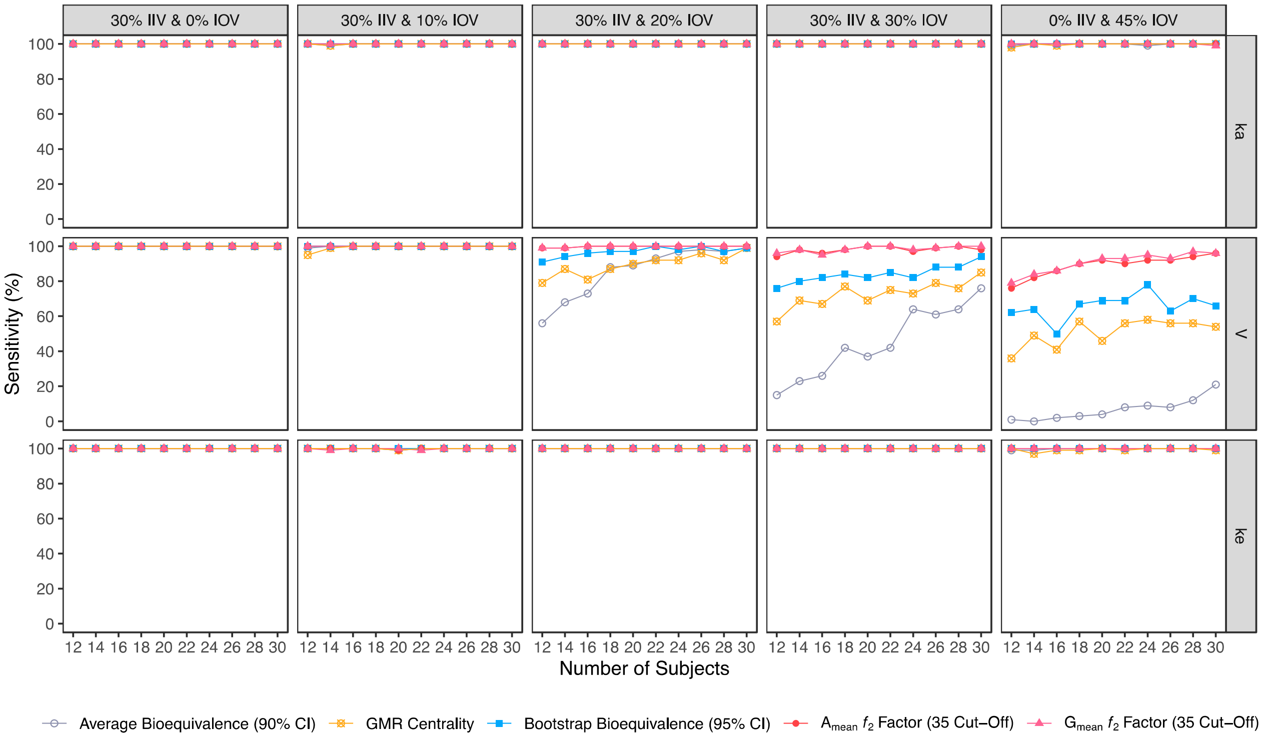

Figure 3. Variation of sensitivity for the bioequivalence evaluation methods (average bioequivalence, centrality of the test-to-reference GMR, bootstrap bioequivalence analysis, and A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35) as function of the number of subjects, per tested variability for the different pharmacokinetic model parameters.

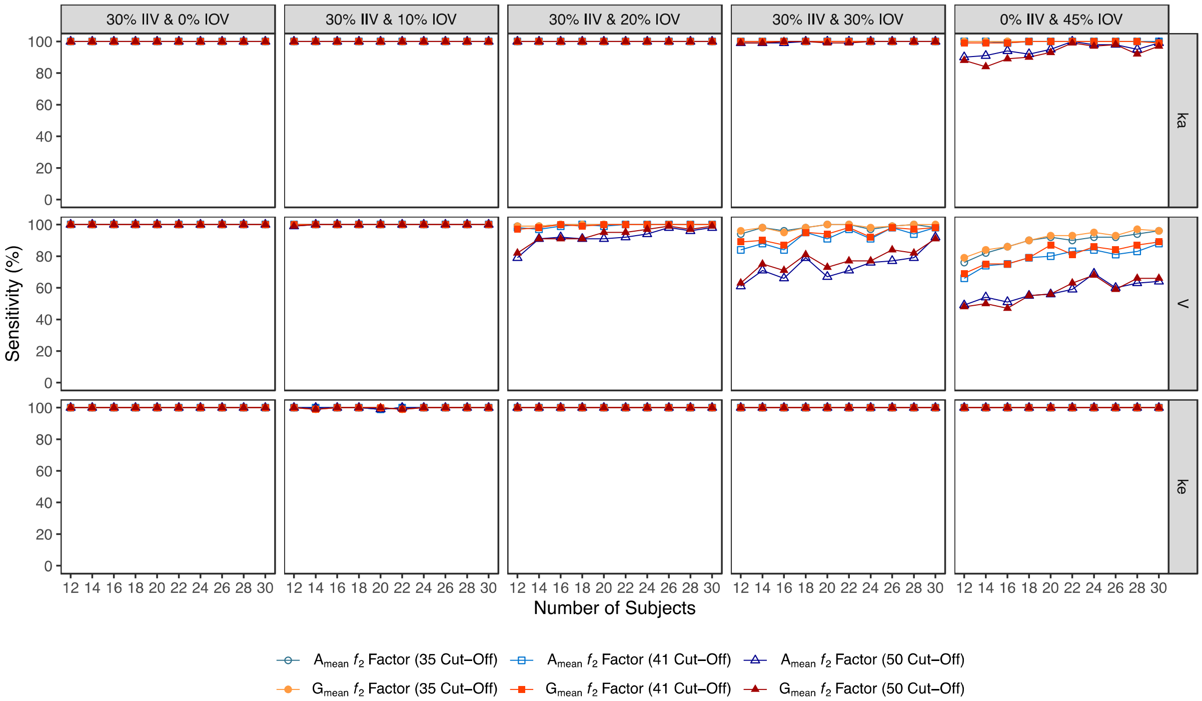

Figure 4. Variation of sensitivity for A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35, 41, and 50, as function of the number of subjects, per tested variability for the different pharmacokinetic model parameters.

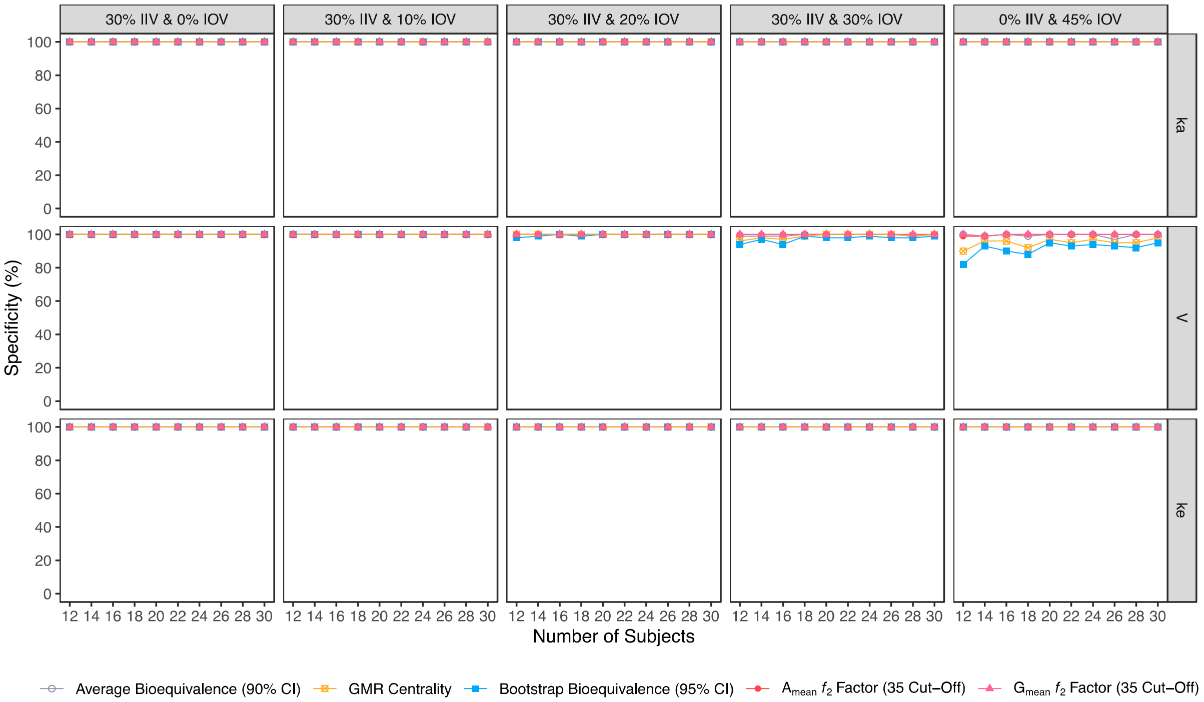

Figure 5. Variation of specificity for the bioequivalence evaluation methods (average bioequivalence, centrality of the test-to-reference GMR, bootstrap bioequivalence analysis, and A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35) as function of the number of subjects, per tested variability for the different pharmacokinetic model parameters.

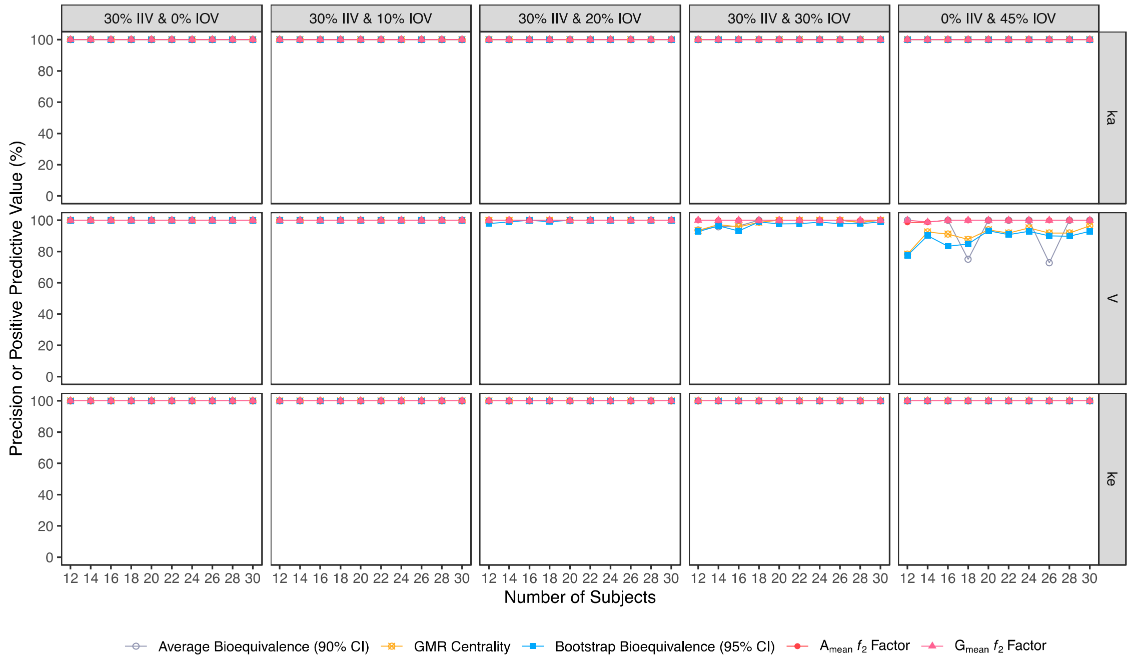

Figure 6. Variation of precision for the bioequivalence evaluation methods (average bioequivalence, centrality of the test-to-reference GMR, bootstrap bioequivalence analysis, and A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35) as function of the number of subjects, per tested variability for the different pharmacokinetic model parameters.

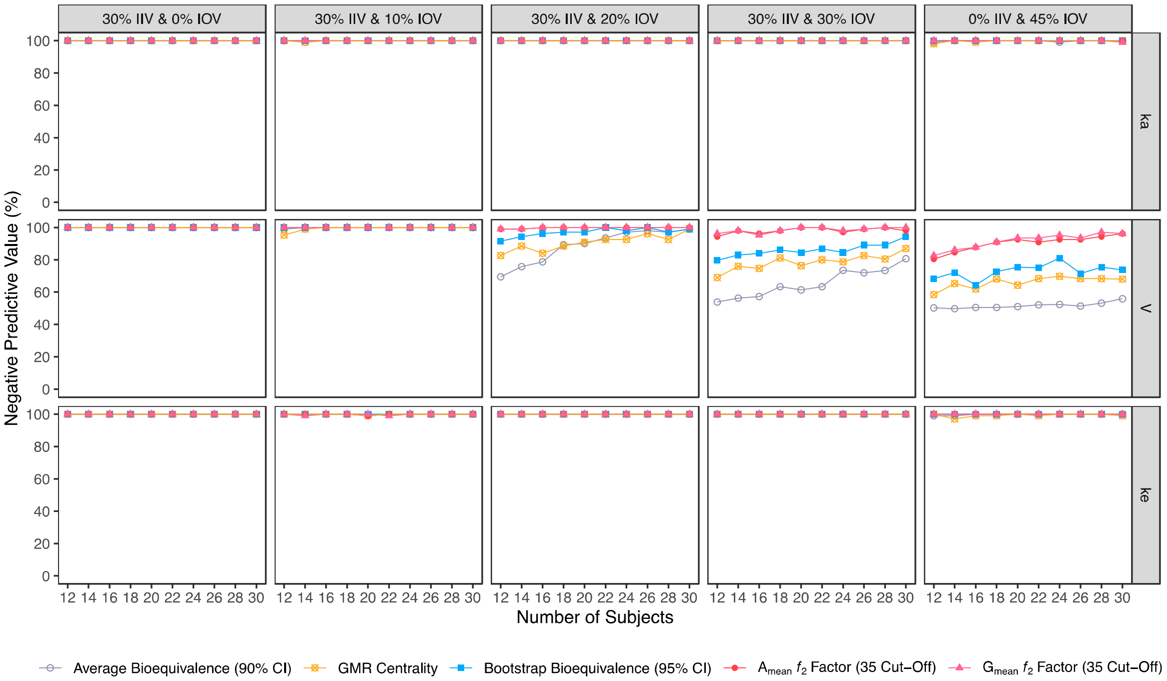

Figure 7. Variation of negative predictive value (NPV) for the bioequivalence evaluation methods (average bioequivalence, centrality of the test-to-reference GMR, bootstrap bioequivalence analysis, and A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35) as function of the number of subjects, per tested variability for the different pharmacokinetic model parameters.

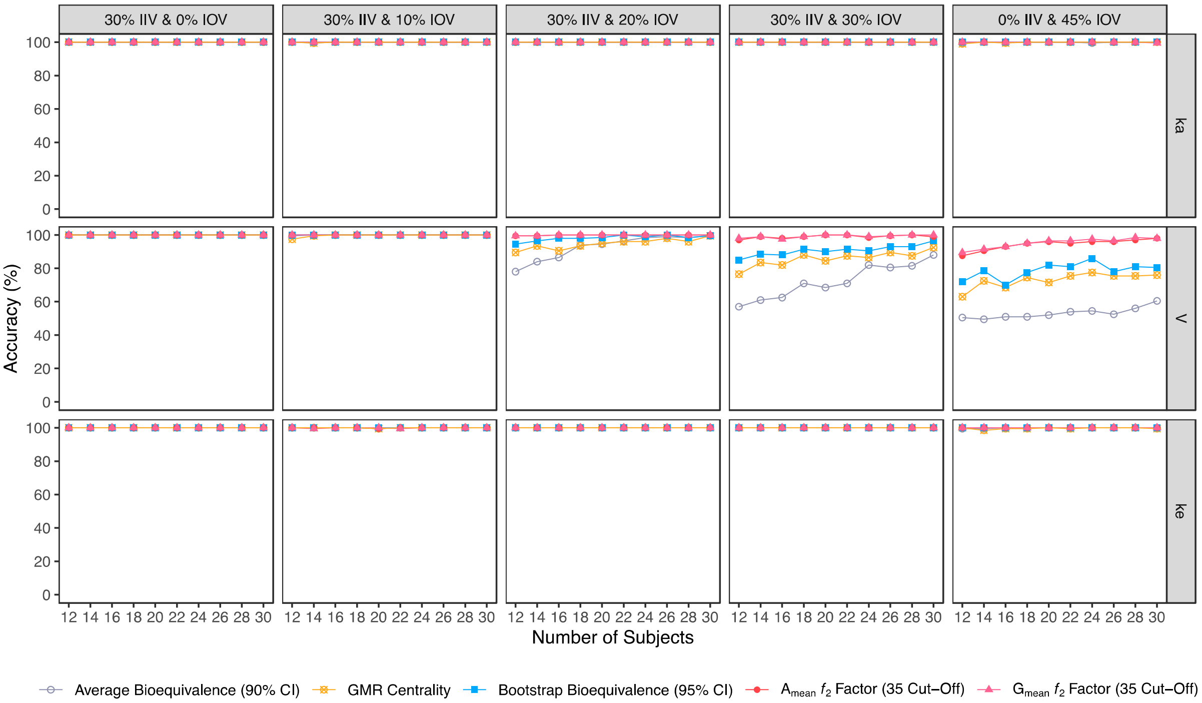

Figure 8. Variation of accuracy for the bioequivalence evaluation methods (average bioequivalence, centrality of the test-to-reference GMR, bootstrap bioequivalence analysis, and A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35) as a function of the number of subjects, per tested variability for the different pharmacokinetic model parameters.

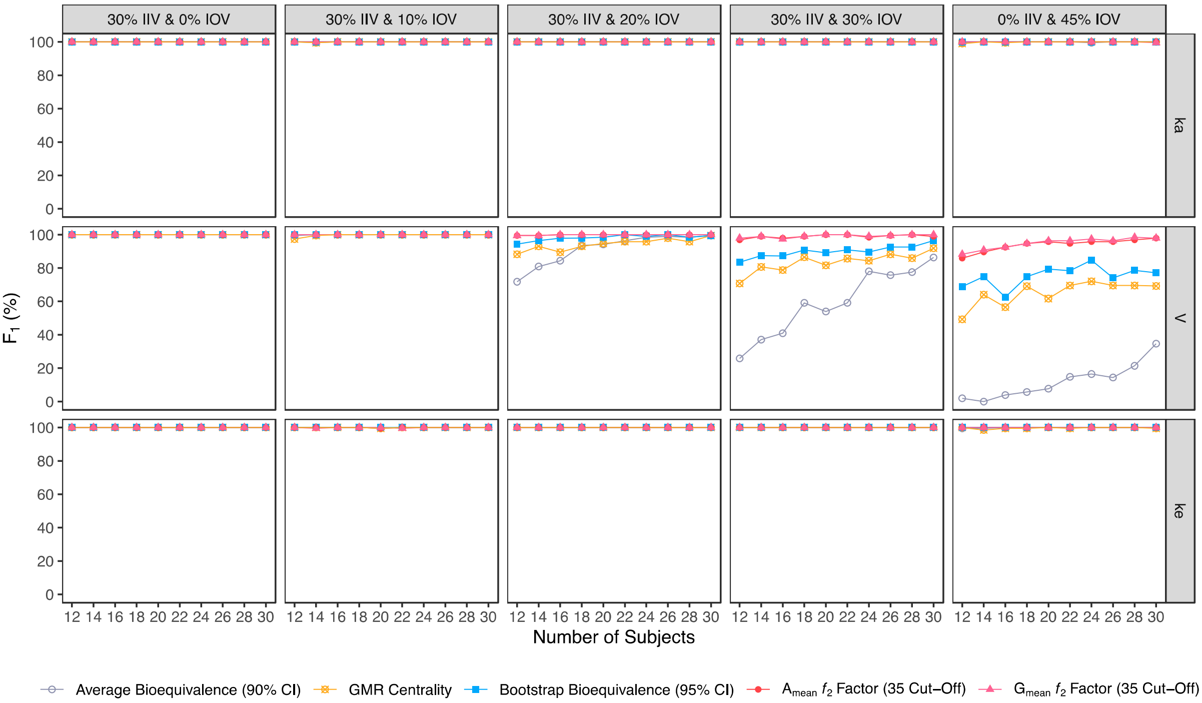

Figure 9. Variation of F_1 for the bioequivalence evaluation methods (average bioequivalence, centrality of the test-to-reference GMR, bootstrap bioequivalence analysis, and A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35) as a function of the number of subjects, per tested variability for the different pharmacokinetic model parameters.

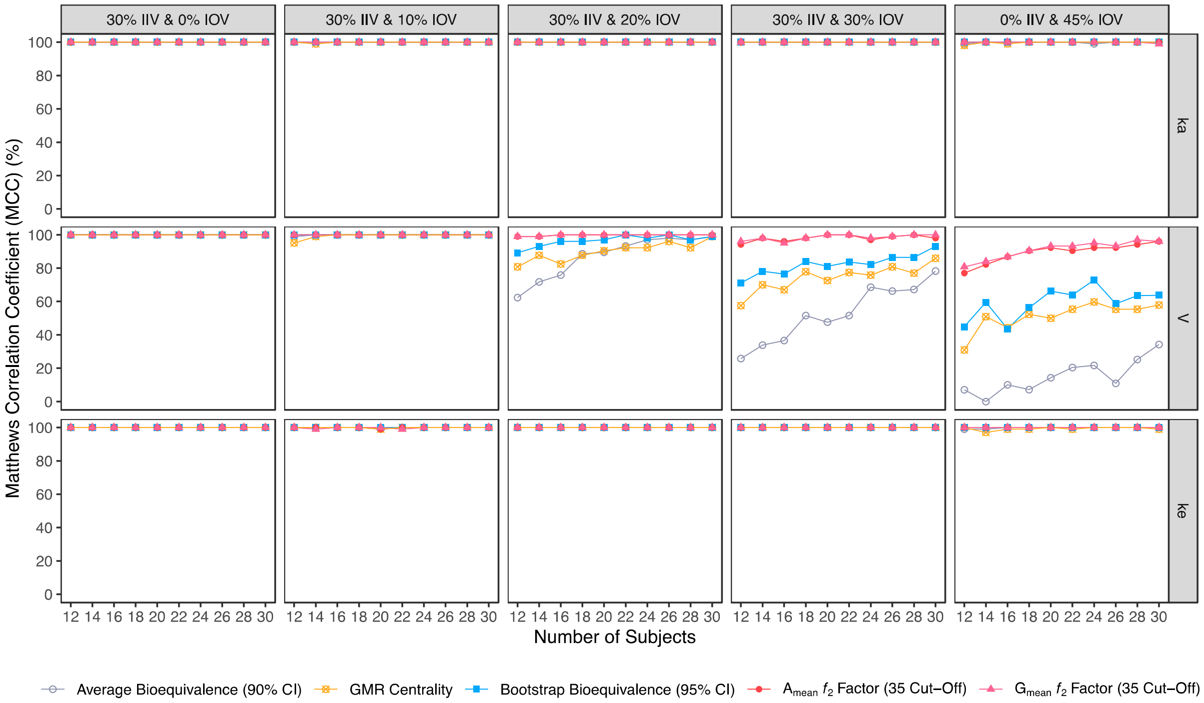

Figure 10. Variation of Matthews correlation coefficient (MCC) for the bioequivalence evaluation methods (average bioequivalence, centrality of the test-to-reference GMR, bootstrap bioequivalence analysis, and A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35) as function of the number of subjects, per tested variability for the different pharmacokinetic model parameters.

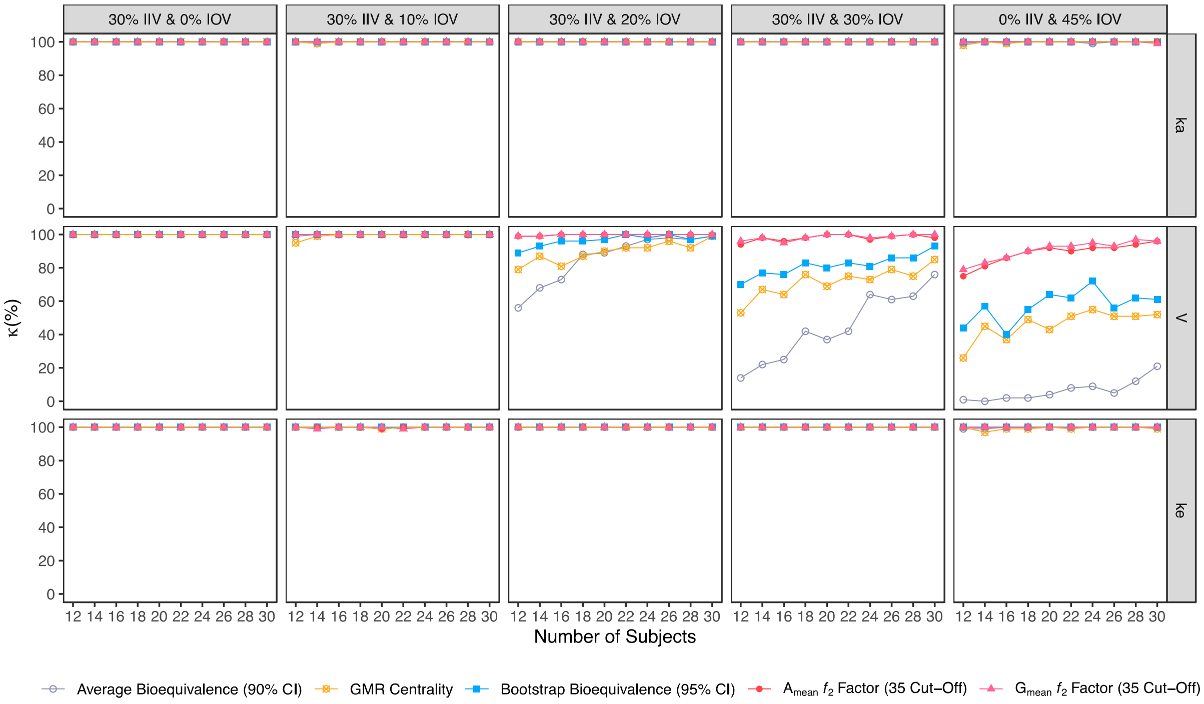

Figure 11. Variation of Cohen's κ for the bioequivalence evaluation methods (average bioequivalence, centrality of the test-to-reference GMR, bootstrap bioequivalence analysis, and A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35) as a function of the number of subjects, per tested variability for the different pharmacokinetic model parameters.

3.2.1. Average Bioequivalence Method

For an IOV of 20% in V and based on the sensitivity evaluation in studies with 12 subjects, the average bioequivalence method correctly identified 56% of the truly bioequivalent test formulations. When incrementing the sample size for 30 subjects, the sensitivity of the average bioequivalence method increased to 99%. When IOV increased to 30%, the sensitivity of the average bioequivalence method decreased to 15% and 76% for studies with 12 and 30 subjects, respectively. For an IOV of 45%, the sensitivity decreased to 1.0% and 21% for studies with 12 and 30 subjects, respectively (Table 3, Figure 3) Nevertheless, the average bioequivalence method performed well avoiding type I errors, even on the highest level of variability, with a specificity rounding the 100% (Table 3, Figure 5).

3.2.2. Centrality of the Test-to-Reference GMR Method

For the simulated scenarios, the centrality of the point estimate (within [90.00–111.11]%) derived from the average bioequivalence approach showed a higher sensitivity than the corresponding 90% CI (Table 4, Figure 3). For a 20% IOV in V, this method correctly identified 79% and 99% of the truly bioequivalent test formulations to be within [90.00–111.11]% in studies with 12 subjects and with 30 subjects, respectively. When increasing IOV to 30%, the sensitivity of the method decreased to 57% and 85% in studies with 12 and 30 subjects, respectively. For an IOV of 45%, the sensitivity of the method decreased to 36% and 54% in studies with 12 and 30 subjects, respectively (Table 4, Figure 3). In terms of specificity, for an IOV of 45%, the centrality of the point estimate method led to an inflation of type I error to 10% in studies simulated with 12 subjects, which can be minimized by increasing the number of subjects in the trials (Table 4, Figure 5).

3.2.3. Bootstrap Bioequivalence Method

The bootstrap bioequivalence method showed a higher sensitivity than the standard parametric approach. For an IOV of 20% in V, this non-parametric method correctly identified more than 90% of truly bioequivalent formulations, irrespective of sample size. When increasing IOV to 30%, the sensitivity decreased to 76% in studies with 12 subjects but scored higher than 90% in studies with 30 subjects (Table 5, Figure 3). For the highest tested variability (IOV of 45%), bootstrap sensitivity was 62% in studies with 12 subjects and 66% in studies with 30 subjects. On the other hand, bootstrap was the method that induced most type I errors (Table 5, Figure 5).

3.2.4. Similarity f_2 Factor Method

The f_2 factor derived from the A_{mean} and G_{mean} pharmacokinetic profiles behaved similarly. For an IOV of 20% in V, and using a cut-off of 35 (i.e., to detect a mean difference of 20%), the f_2 method could correctly identify more than 99% of truly bioequivalent test formulations in studies with 12 and 30 subjects. When increasing IOV to 30%, the sensitivity slightly decreased to 94% in studies with 12 subjects but scored higher than 98% in studies with 30 subjects. For the highest tested variability (IOV of 45%), the f_2 factor derived from both A_{mean} and G_{mean} profiles was found to be a much more sensitive approach than the standard average bioequivalence approach, with >76% and 96% of truly bioequivalent test formulations identified in studies with 12 and 30 subjects, respectively (Tables 6 and 7, and Figure 3).

Using a cut-off of 41 (i.e., to detect a mean difference of 15%), the f_2 factor method still performed better than the average bioavailability method. As expected, the sensitivity slightly decreased while using a higher cut-off, however, differences in the sensitivity between 35 and 41 cut-off values were only noticeable at 30% IOV. For an IOV of 30% in V, the f_2 method could correctly identify more than 84% of truly bioequivalent test formulations with 12 subjects (nearly a 10% decrease in comparison to a cut-off of 35) and 98% with 30 subjects (no difference between the two cut-off values). Moreover, for the highest tested variability (IOV of 45%), the sensitivity of f_2 factor method for both A_{mean} and G_{mean} profiles using a cut-off of 41 decreased to nearly 66% and 88% (nearly 10% decrease in comparison to a cut-off of 35) with 12 and 30 subjects, respectively (Tables 8 and 9, and Figure 3).

Using a cut-off of 50 (i.e., to detect a mean difference of 10%), f_2 factor method performed slightly worse than the average bioequivalence method in studies simulated for the lowest sample size (12 subjects) with the highest variability (IOV of 45%) on k_a. In this case, a sensitivity of 88% was attained. However, regarding the different variability scenarios in V, the f_2 factor method using a cut-off of 50 was always more sensitive than the average bioequivalence method for an IOV $\geq$ 20%. Nevertheless, as expected, the sensitivity decreased, compared to the other tested cut-offs. For an IOV of 20% in V and using a cut-off value of 50, f_2 factor method correctly predicted nearly 80% of the truly bioequivalent test formulations with only 12 subjects and 99% with 30 subjects. For an IOV of 30%, f_2 factor method showed a sensitivity of more than 60% in studies with 12 subjects and more than 90% with 30 subjects. For the highest tested variability (IOV of 45%), f_2 factor correctly predicted almost 50% of the truly bioequivalent test formulations with 12 subjects and more than 64% with 30 subjects (Tables 10 and 11, and Figure 3).

Along with the higher sensitivity shown by the f_2 factor method using different cut-offs in comparison to the average bioequivalence method, no inflation of type I error (>5%) was induced with f_2 factor method for A_{mean} and G_{mean} pharmacokinetic profiles, using all cut-off values (Tables 5–11, Figure 5).

3.2.5. Comparison of Average Bioequivalence, Centrality of the Point Estimate, Bootstrap Bioequivalence, and Similarity f_2 Factor Methods

Accuracy, MCC, F_1 and κ were calculated in order to select the best methodology to assess the potential of a test formulation to be bioequivalent to a reference formulation on the rate of drug absorption, based on pilot BA/BE trials.

In general, average bioequivalence was the least accurate approach. For an IOV of 20% in V and for a minimum sample size of 12 subjects, the average bioequivalence method showed an accuracy of 78% (Table 3), while the other approaches scored $\geq$89.5% (Figure 8). When increasing the sample size to 30 subjects, all methods were $\geq$99% accurate. By increasing IOV to 30% in V, the accuracy of the average bioequivalence approach decreased to 57% and 88% with 12 and 30 subjects, respectively (Table 3); and the accuracy of the centrality of the GMR approach scored between 76.5 and 92.5% with 12 and 30 subjects, respectively (Table 4). All the other methods scored similarly, with an accuracy above 80% for studies with 12 subjects and above 95% for studies with 30 subjects (Figure 8). At the highest tested level of variability (IOV of 45%), the accuracy of the average bioequivalence method decreased to 50.5% and 60.5% with sample sizes of 12 and 30 subjects, respectively (Table 3); and the accuracy of the centrality of the point estimate method was decreased to 63% and 76% for studies with 12 and 30 subjects, respectively (Table 4). The bootstrap bioequivalence method showed an accuracy of 72% and 80.5% in studies with 12 and 30 subjects, respectively (Table 5). Regarding the f_2 factor method derived from A_{mean} and G_{mean} pharmacokinetic profiles, the accuracy was above 80% and 94% in studies with 12 and 30 subjects, respectively, using a cut-off of 35 (Tables 6 and 7) and a cut-off of 41 (Tables 8 and 9); while using a cut-off of 50, accuracy ranged from 74% to 82% in studies with 12 and 30 subjects, respectively (Tables 10 and 11).

Similarly, average bioequivalence was the method with the lowest harmonic mean between sensitivity and precision (F_1) (Figure 9). For an IOV of 20% in V, and for studies with 12 and 30 subjects, F_1 estimates for the average bioequivalence ranged between 71.8% and 99.5%, respectively (Table 3); for the centrality of the GMR ranged between 88.3% and 99.5% (Table 4); and for bootstrap bioequivalence ranged between 94.3% and 99.5% (Table 5). For the f_2 factor derived from A_{mean} and G_{mean} pharmacokinetic profiles, F_1 was above 99% using a cut-off of 35 (Tables 6 and 7), above 98% using a cut-off and 41 (Tables 8 and 9), while using a cut-off of 50, F_1 ranged from 88% to 99% in studies with 12 and 30 subjects, respectively (Tables 10 and 11).

Table 3. Cross-Tabulated Matrix Statistics Calculated for Average Bioequivalence (90% CI).

	Average Bioequivalence (90% CI)									
	Sensitivity (%)	**Type II Error (%)**	**Specificity (%)**	**Type I Error (%)**	**Precision (%)**	**NPV (%)**	**Accuracy (%)**	**F_1 (%)**	**MCC (%)**	**κ (%)**
Baseline	100	0.00	100	0.00	100	100	100	100	100	100
k_a										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	99.0–100	1.00–0.00	100	0.00	100	99.0–100	99.5–100	99.5–100	99.0–100.0	99.0–100
V										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	99.0–100	1.00–0.00	100	0.00	100	99.0–100	99.5–100	99.5–100	99.0–100	99.0–100
30% IIV & 20% IOV	56.0–99.0	44.0–1.00	100	0.00	100	69.4–99.0	78.0–99.5	71.8–99.5	62.4–99.0	56.0–99.0
30% IIV & 30% IOV	15.0–76.0	85.0–24.0	99.0–100	1.00–0.00	93.75–100	53.8–80.6	57.0–88.0	25.9–86.4	25.8–78.3	14.0–76.0
0% IIV & 45% IOV	1.00–21.0	99.0–79.0	100	0.00	100	50.3–55.9	50.5–60.5	1.98–34.7	7.09–34.3	1.00–21.0
k_e										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	99.0–100	1.00–0.00	100	0.00	100	99.0–100	99.5–100	99.5–100	99.0–100	99.0–100

Values represent the range calculated from simulated studies with 12 and 30 subjects. When statistics do not change between 12 and 30 subjects, unique values are presented instead of ranges. F_1—Harmonic mean of sensitivity and precision; κ—Cohen's Kappa; MCC—Matthews correlation coefficient; NPV—Negative predictive value.

Table 4. Cross-Tabulated Matrix Statistics Calculated for test-to-reference GMR Centrality.

	Test-to-Reference GMR Centrality									
	Sensitivity (%)	**Type II Error (%)**	**Specificity (%)**	**Type I Error (%)**	**Precision (%)**	**NPV (%)**	**Accuracy (%)**	**F_1 (%)**	**MCC (%)**	**κ (%)**
Baseline	100	0.00	100	0.00	100	100	100	100	100	100
k_a										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	98.0–100	2.00–0.00	100	0.00	100	98.0–100	99.0–100	99.0–100	98.02–100	98.0–100
V										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	95.0–100	5.00–0.00	100	0.00	100	95.2–100	97.5–100	97.4–100	95.1–100	95.0–100
30% IIV & 20% IOV	79.0–99.0	21.0–1.00	100	0.00	100	82.6–99.0	89.5–99.5	88.3–99.5	80.8–99.0	79.0–99.0
30% IIV & 30% IOV	57.0–85.0	43.0–15.0	96.0–100	4.00–0.00	93.4–100	69.1–87.0	76.5–92.5	70.8–91.9	57.6–86.0	53.0–85.0
0% IIV & 45% IOV	36.0–54.0	64.0–46.0	90.0–98.0	10.0–2.00	78.3–96.4	58.4–68.1	63.0–76.0	49.3–69.2	30.9–57.9	26.0–52.0
k_e										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	99.0–100	1.00–0.00	100	0.00	100	99.0–100	99.5–100	99.5–100	99.0–100	99.0–100

Values represent the range calculated from simulated studies with 12 and 30 subjects. When statistics do not change between 12 and 30 subjects, unique values are presented instead of ranges. F_1—Harmonic mean of sensitivity and precision; κ—Cohen’s Kappa; MCC—Matthews correlation coefficient; NPV—Negative predictive value.

Table 5. Cross-Tabulated Matrix Statistics Calculated for Bootstrap Bioequivalence (95% CI).

	Bootstrap Bioequivalence (95% CI)									
	Sensitivity (%)	**Type II Error (%)**	**Specificity (%)**	**Type I Error (%)**	**Precision (%)**	**NPV (%)**	**Accuracy (%)**	**F_1 (%)**	**MCC (%)**	**κ (%)**
Baseline	100	0.00	100	0.00	100	100	100	100	100	100
k_a										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100
V										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	91.0–99.0	9.00–1.00	98.0–100	2.00–0.00	97.85–100	91.6–99.0	94.5–99.5	94.3–99.5	89.2–99.0	89.0–99.0
30% IIV & 30% IOV	76.0–94.0	24.0–6.00	94.0–99.0	6.00–1.00	92.7–98.9	79.7–94.3	85.0–96.5	83.5–96.4	71.2–93.1	70.0–93.0
0% IIV & 45% IOV	62.0–66.0	38.0–34.0	82.0–95.0	18.0–5.00	77.5–93.0	68.3–73.6	72.0–80.5	68.9–77.2	44.9–63.7	44.0–61.0
k_e										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100

Values represent the range calculated from simulated studies with 12 and 30 subjects. When statistics do not change between 12 and 30 subjects, unique values are presented instead of ranges. F_1—Harmonic mean of sensitivity and precision; κ—Cohen's Kappa; MCC—Matthews correlation coefficient; NPV—Negative predictive value.

Table 6. Cross-Tabulated Matrix Statistics Calculated for Arithmetic Mean (A_{mean}) f_2 Factor, Using a Cut-Off of 35.

	A_{mean} f_2 Factor (Cut-Off of 35)									
	Sensitivity (%)	**Type II Error (%)**	**Specificity (%)**	**Type I Error (%)**	**Precision (%)**	**NPV (%)**	**Accuracy (%)**	**F_1 (%)**	**MCC (%)**	**κ (%)**
Baseline	100	0.00	100	0.00	100	100	100	100	100	100
k_a										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100
V										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	99.0–100	1.00–0.00	100	0.00	100	99.0–100	99.5–100	99.5–100	99.0–100	99.0–100
30% IIV & 30% IOV	94.0–98.0	6.00–2.00	100	0.00	100	94.3–98.0	97.0–99.0	96.9–99.0	94.2–98.0	94.0–98.0
0% IIV & 45% IOV	76.0–96.0	24.0–4.00	99.0–100	1.00–0.00	98.7–100	80.5–96.2	87.5–98.0	85.9–98.0	77.1–96.1	75.0–96.0
k_e										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100

Values represent the range calculated from simulated studies with 12 and 30 subjects. When statistics do not change between 12 and 30 subjects, unique values are presented instead of ranges. F_1—Harmonic mean of sensitivity and precision; κ—Cohen's Kappa; MCC—Matthews correlation coefficient; NPV—Negative predictive value.

Table 7. Cross-Tabulated Matrix Statistics Calculated for Geometric Mean (G_{mean}) f_2 Factor, Using a Cut-Off of 35.

	G_{mean} f_2 Factor (Cut-Off of 35)									
	Sensitivity (%)	**Type II Error (%)**	**Specificity (%)**	**Type I Error (%)**	**Precision (%)**	**NPV (%)**	**Accuracy (%)**	**F_1 (%)**	**MCC (%)**	**κ (%)**
Baseline	100	0.00	100	0.00	100	100	100	100	100	100
k_a										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100
V										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	99.0–100	1.00–0.00	100	0.00	100	99.0–100	99.5–100	99.5–100	99.0–100	99.0–100
30% IIV & 30% IOV	96.0–100	4.00–0.00	100	0.00	100	96.2–100	98.0–100	98.0–100	96.1–100	96.0–100
0% IIV & 45% IOV	79.0–96.0	21.0–4.00	100	0.00	100	82.6–96.2	89.5–98.0	88.3–98.0	80.8–96.1	79.0–96.0
k_e										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100

Values represent the range calculated from simulated studies with 12 and 30 subjects. When statistics do not change between 12 and 30 subjects, unique values are presented instead of ranges. F_1—Harmonic mean of sensitivity and precision; κ—Cohen's Kappa; MCC—Matthews correlation Coefficient; NPV—Negative predictive value.

Table 8. Cross-Tabulated Matrix Statistics Calculated for Arithmetic Mean (A_{mean}) f_2 Factor, Using a Cut-Off of 41.

	A_{mean} f_2 Factor (Cut-Off of 41)									
	Sensitivity (%)	**Type II Error (%)**	**Specificity (%)**	**Type I Error (%)**	**Precision (%)**	**NPV (%)**	**Accuracy (%)**	**F_1 (%)**	**MCC (%)**	**κ (%)**
Baseline	100	0.00	100	0.00	100	100	100	100	100	100
k_a										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100
V										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	98.0–100	2.00–0.00	100	0.00	100	98.0–100	99.0–100	99.0–100	98.0–100	98.0–100
30% IIV & 30% IOV	84.0–98.0	16.0–2.00	100	0.00	100	86.2–98.0	92.0–99.0	91.3–99.0	85.1–98.0	84.0–98.0
0% IIV & 45% IOV	66.0–88.0	34.0–12.0	100	0.00	100	74.6–89.3	83.0–94.0	79.5–93.6	70.2–88.6	66.0–88.0
k_e										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100

Values represent the range calculated from simulated studies with 12 and 30 subjects. When statistics do not change between 12 and 30 subjects, unique values are presented instead of ranges. F_1—Harmonic mean of sensitivity and precision; κ—Cohen's Kappa; MCC—Matthews correlation coefficient; NPV—Negative predictive value.

Table 9. Cross-Tabulated Matrix Statistics Calculated for Geometric Mean (G_{mean}) f_2 Factor, Using a Cut-Off of 41.

	G_{mean} f_2 Factor (Cut-Off of 41)									
	Sensitivity (%)	**Type II Error (%)**	**Specificity (%)**	**Type I Error (%)**	**Precision (%)**	**NPV (%)**	**Accuracy (%)**	**F_1 (%)**	**MCC (%)**	**κ (%)**
Baseline	100	0.00	100	0.00	100	100	100	100	100	100
k_a										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	99.0	1.00	100	0.00	100	99.0	99.5	99.5	99.0	99.0
V										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	97.0–100	3.00–0.00	100	0.00	100	97.1–100	98.5–100	98.5–100	97.0–100	97.0–100
30% IIV & 30% IOV	89.0–98.0	11.0–2.00	100	0.00	100	90.1–98.0	94.5–99.0	94.2–99.0	89.5–98.0	89.0–98.0
0% IIV & 45% IOV	69.0–89.0	31.0–11.0	100	0.00	100	76.3–90.1	84.5–94.5	81.7–94.2	72.6–89.5	69.0–89.0
k_e										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100

Values represent the range calculated from simulated studies with 12 and 30 subjects. When statistics do not change between 12 and 30 subjects, unique values are presented instead of ranges. F_1—Harmonic mean of sensitivity and precision; κ—Cohen's Kappa; MCC—Matthews correlation coefficient; NPV—Negative predictive value.

Table 10. Cross-Tabulated Matrix Statistics Calculated for Arithmetic Mean (A_{mean}) f_2 Factor, Using a Cut-Off of 50.

	A_{mean} f_2 Factor (Cut-Off of 50)									
	Sensitivity (%)	**Type II Error (%)**	**Specificity (%)**	**Type I Error (%)**	**Precision (%)**	**NPV (%)**	**Accuracy (%)**	**F_1 (%)**	**MCC (%)**	**κ (%)**
Baseline	100	0.00	100	0.00	100	100	100	100	100	100
k_a										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	99.0–100	1.00–0.00	100	0.00	100	99.0–100	99.5–100	99.5–100	99.0–100	99.0–100
0% IIV & 45% IOV	90.0–99.0	10.0–1.00	100	0.00	100	90.9–99.0	95.0–99.5	94.7–99.5	90.5–99.0	90.0–99.0
V										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	99.0–100	1.00–0.00	100	0.00	100	99.0–100	99.5–100	99.5–100	99.0–100	99.0–100
30% IIV & 20% IOV	79.0–98.0	21.0–2.00	100	0.00	100	82.6–98.0	89.5–99.0	88.3–99.0	80.8–98.0	79.0–98.0
30% IIV & 30% IOV	61.0–92.0	39.0–8.00	100	0.00	100	71.9–92.6	80.5–96.0	75.8–95.8	66.3–92.3	61.0–92.0
0% IIV & 45% IOV	49.0–64.0	51.0–36.0	100	0.00	100	66.2–73.5	74.5–82.0	65.8–78.1	57.0–68.6	49.0–64.0
k_e										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100

Values represent the range calculated from simulated studies with 12 and 30 subjects. When statistics do not change between 12 and 30 subjects, unique values are presented instead of ranges. F_1—Harmonic mean of sensitivity and precision; κ—Cohen's Kappa; MCC—Matthews correlation coefficient; NPV—Negative predictive value.

Table 11. Cross-Tabulated Matrix Statistics Calculated for Geometric Mean (G_{mean}) f_2 Factor, Using a Cut-Off of 50.

	G_{mean} f_2 Factor (Cut-Off of 50)									
	Sensitivity (%)	**Type II Error (%)**	**Specificity (%)**	**Type I Error (%)**	**Precision (%)**	**NPV (%)**	**Accuracy (%)**	**F_1 (%)**	**MCC (%)**	**κ (%)**
Baseline	100	0.00	100	0.00	100	100	100	100	100	100
k_a										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	99.0–100	1.00–0.00	100	0.00	100	99.0–100	99.5–100	99.5–100	99.0–100	99.0–100
% IIV & 45% IOV	88.0–97.0	12.0–3.00	100	0.00	100	89.3–97.1	94.0–98.5	93.6–98.5	88.6–97.0	88.0–97.0
V										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	99.0–100	1.00–0.00	100	0.00	100	99.0–100	99.5–100	99.5–100	99.0–100	99.0–100
30% IIV & 20% IOV	82.0–99.0	18.0–1.00	100	0.00	100	84.8–99.0	91.0–99.5	90.1–99.5	83.4–99.0	82.0–99.0
30% IIV & 30% IOV	63.0–91.0	37.0–9.00	100	0.00	100	73.0–91.7	81.5–95.5	77.3–95.3	67.8–91.4	63.0–91.0
0% IIV & 45% IOV	48.0–66.0	52.0–34.0	100	0.00	100	65.8–74.6	74.0–83.0	64.9–79.5	56.2–70.2	48.0–66.0
k_e										
30% IIV & 0% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 10% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 20% IOV	100	0.00	100	0.00	100	100	100	100	100	100
30% IIV & 30% IOV	100	0.00	100	0.00	100	100	100	100	100	100
0% IIV & 45% IOV	100	0.00	100	0.00	100	100	100	100	100	100

Values represent the range calculated from simulated studies with 12 and 30 subjects. When statistics do not change between 12 and 30 subjects, unique values are presented instead of ranges. F_1—Harmonic mean of sensitivity and precision; κ—Cohen's Kappa; MCC—Matthews correlation coefficient; NPV—Negative predictive value.

For an IOV of 30% and for studies with 12 and 30 subjects, average bioequivalence F_1 highly decreased, ranging between 25.9% and 86.4%, respectively (Table 3). For the same sample sizes, the centrality of the GMR method showed an F_1 between 70.8% and 91.9% (Table 4), respectively, and the bootstrap bioequivalence method presented an F_1 between 83.5% and 96.4% (Table 5). For the f_2 factor derived from A_{mean} and G_{mean} pharmacokinetic profiles, F_1 was above 96% using a cut-off of 35 (Tables 6 and 7), and above 91% using a cut-off of 41 (Tables 8 and 9). Using a cut-off of 50, F_1 ranged from 75% to 95% in studies with 12 and 30 subjects, respectively (Tables 10 and 11).

For the highest IOV (45%) in V, and for studies with 12 and 30 subjects, average bioequivalence F_1 decreased to 2% and 34.7%, respectively (Table 3). For the same sample sizes, the centrality of the GMR showed an F_1 between 49.3% and 69.2% (Table 4), respectively, and the bootstrap bioequivalence method presented an F_1 of 68.9% and 77.2% (Table 5). For the f_2 factor derived from A_{mean} and G_{mean} pharmacokinetic profiles, F_1 ranged within 85% and 98% using a cut-off of 35 (Tables 6 and 7), within 80% and 94% using a cut-off and 41 (Tables 8 and 9); while using a cut-off of 50, F_1 ranged from 65% to 78% in studies with 12 and 30 subjects, respectively (Tables 10 and 11).

Considering the correlation between the true classes and the predicted labels (MCC) (Figure 10), average bioequivalence was the method that scored lower. For an IOV of 20% in V, average bioequivalence MCC ranged between 62.4% and 99% in studies with 12 and 30 subjects, respectively (Table 3). For the same sample sizes, MCC for the centrality of the point estimate ranged between 80.8% and 99.0% (Table 4) and MCC for the bootstrap bioequivalence ranged between 89.2% and 99% (Table 5). For the f_2 factor derived from A_{mean} and G_{mean} pharmacokinetic profiles, MCC ranged within 99% and 100% using a cut-off of 35 (Tables 6 and 7), within 97% and 100% using a cut-off and 41 (Tables 8 and 9); while using a cut-off of 50, MCC ranged from 80% to 98% in studies with 12 and 30 subjects, respectively (Tables 10 and 11).

For an IOV of 30%, average bioequivalence MCC ranged between 25.8% and 78.3% in studies with 12 and 30 subjects, respectively (Table 3). For the same sample sizes, MCC for the centrality of the point estimate ranged between 57.6% and 86.0% (Table 4) and MCC for the bootstrap bioequivalence ranged between 71.2% and 93.1% (Table 5). For the f_2 factor derived from A_{mean} and G_{mean} pharmacokinetic profiles, MCC ranged within 94% and 100% using a cut-off of 35 (Tables 6 and 7), within 85% and 98% using a cut-off and 41 (Tables 8 and 9); while using a cut-off of 50, MCC ranged from 66% to 91% in studies with 12 and 30 subjects, respectively (Tables 10 and 11).

For the highest IOV (45%) in V, average bioequivalence MCC decreased to 7.10% and 34.3% in studies with 12 and 30 subjects, respectively (Table 3). For the same sample sizes, MCC for the centrality of the point estimate ranged between 30.9% and 57.9% (Table 4) and MCC for the bootstrap bioequivalence ranged between 44.9% and 63.7% (Table 5). For the f_2 factor derived from A_{mean} and G_{mean} pharmacokinetic profiles, MCC ranged between 77% and 96% using a cut-off of 35 (Tables 6 and 7), within 70% and 89% using a cut-off and 41 (Tables 8 and 9); while using a cut-off of 50, MCC ranged from 56% to 70% in studies with 12 and 30 subjects, respectively (Tables 10 and 11).

Average bioequivalence was the method with the lowest concordance agreement relative to what would be expected by chance (κ) (Figure 11). For an IOV of 20% in V, average bioequivalence κ ranged between 56% and 99% in studies with 12 and 30 subjects, respectively (Table 3); For the same sample sizes, κ for the centrality of the point estimate ranged between 79.0% and 100% (Table 4) and κ for bootstrap bioequivalence ranged between 89.0% and 99.0% (Table 5). For the f_2 factor derived from A_{mean} and G_{mean} pharmacokinetic profiles, κ was above 99% using a cut-off of 35 (Tables 6 and 7), above 97% using a cut-off and 41 (Tables 8 and 9); while using a cut-off of 50, κ ranged from 79% to 99% in studies with 12 and 30 subjects, respectively (Tables 10 and 11).

For an IOV of 30%, average bioequivalence κ ranged between 14% and 76% in studies with 12 and 30 subjects, respectively (Table 3). For the same sample sizes, the centrality of the point estimate ranged between 53% and 85% (Table 4) and κ for the bootstrap bioequivalence ranged between 70% and 93% (Table 5). For the f_2 factor derived from A_{mean} and G_{mean} pharmacokinetic profiles, κ was above 94% using a cut-off of 35 (Tables 6 and 7), above 84% using a cut-off and 41 (Tables 8 and 9); while using a cut-off of 50, κ ranged from 61% to 92% in studies with 12 and 30 subjects, respectively (Tables 10 and 11).

For the highest IOV (45%) in V, average bioequivalence κ decreased to 1% and 21% in studies with 12 and 30 subjects, respectively (Table 3). For the same sample sizes, κ for the centrality of the point estimate ranged between 26.0% and 52.0% (Table 4) and κ for bootstrap bioequivalence ranged between 44% and 61% (Table 5). For the f_2 factor derived from A_{mean} and G_{mean} pharmacokinetic profiles, κ was within 75% and 96% using a cut-off of 35 (Tables 6 and 7), within 66% and 89% using a cut-off and 41 (Tables 8 and 9); while using a cut-off of 50, κ ranged from 48% to 66% in studies with 12 and 30 subjects, respectively (Tables 10 and 11).

4. Discussion

For the pharmacokinetic model tested, an increment in the variability of V was associated with a higher dispersion of C_{max} values (GCV% ≈ 30–48%). Hence, the within-individual variability (IOV) in V was the identified variability with the highest impact on the bioequivalence evaluation of C_{max}.

For each bioequivalence evaluation method (average bioequivalence, centrality of the test-to-reference GMR, bootstrap bioequivalence, and A_{mean} and G_{mean} f_2 factor evaluated with a cut-off of 35, 41, and 50) the relationship between type I and type II errors was studied. Moreover, accuracy, MCC, F_1, and κ were calculated in order to select the best methodology for the evaluation of the potentiality of a test formulation to be bioequivalent to a reference formulation on the rate of drug absorption, based on pilot BA/BE trials. For each bioequivalence evaluation method, results were consistent for all the calculated cross-tabulation matrix statistics.

Average bioequivalence was found to be the most underpower method tested, i.e., that induced higher type II errors (Table 3 and Figure 3). A critical decrease in sensitivity was observed for an IOV of 20% (in V), a level of variability lower than the cut-off value of 30% established by the EMA's Guideline on the Evaluation of Bioequivalence for highly variable drug/drug products [1]. Similarly, average bioequivalence was the method that scored lower for all the other performance measures. Nevertheless, inflation of the type I error was not observed for this statistical method, being kept below 0.05 (Table 3 and Figure 5).

The newly proposed approaches (centrality of the test-to-reference GMR, bootstrap bioequivalence, and A_{mean} and G_{mean} f_2 factor) showed a higher sensitivity/power than the established average bioequivalence method commonly used (Figure 3).

The alternative methodologies can maintain a power of at least 80% with less than 20 subjects in studies with a high IOV (30%), while the average bioequivalence approach required at least 80 subjects to maintain the same power level (Figure 3, Table 12).

Table 12. Sample Size for a 2 × 2 × 2 Crossover Study for Different Bioequivalence Evaluation Methods, Targeting a Power of at Least 80%, an α of 0.05, and Assuming a GMR of 100%.

IOV (%)	Average Bioequivalence [1]	Bootstrap Bioequivalence	G_{mean} f_2 Factor		
			35	41	50
10%	12	12	12	12	12
20%	16	12	12	12	12
30%	32	14	12	12	18
45%	66	>30	14	20	>30

[1] Calculated using R package 'PowerTOST' version 1.5–3 [14]. G_{mean}—Geometric mean; IOV—Inter-occasion variability.

Moreover, the alternative methods showed a higher performance for the other cross-tabulated matrix statistics, i.e., a better concordance between the truth and predictions.

Hence, for downsized trials as pilot studies, the use of the proposed alternative approaches may reduce the uncertainty in the evaluation of the potentiality of a test formulation to be bioequivalent to a reference formulation on the rate of drug absorption, helping pharmaceutical companies on the decision to go forward to pivotal bioequivalence studies.

Regarding the centrality of the GMR, the method showed a higher sensitivity than the average bioequivalence method. However, using this alternative can be misleading, as it may lead to false positives due to its lower specificity for higher variabilities.

The bootstrap methodology was able to maintain a power of at least 80% for simulations with an IOV of 20% in V and a sample size of 12 subjects (Table 5 and Figure 3), as well as for simulations with an IOV of 30% and a sample size of only 14 subjects (Table 12). These sample sizes correspond to 75% and 44% of the sample size estimated based on the same assumptions of IOV (20% and 30%) and expected power level (80%), using the average bioequivalence analysis approach (i.e., 16 and 32 subjects, respectively, Table 12). However, this non-parametric approach was the method that induced higher type I rates (Table 5 and Figure 5). Nevertheless, the bootstrap bioequivalence method was found to be more accurate than the standard average bioequivalence.

Additionally, f_2 factor methodology was tested using cut-offs of 35, 41, and 50 for testing a mean difference of 20%, 15%, and 10%, respectively, between the concentration-time profiles of test and reference, until the reference C_{max}.

Regarding the f_2 factor methodology, for an IOV of 20% (in V), 12 subjects are needed to target a power of at least 80%, either using a cut-off of 35 or 41, corresponding to 75% of the required sample size estimated with the same IOV and power assumptions, using the average bioequivalence analysis approach (Table 12). Using a cut-off of 50, 14 subjects would be needed (corresponding to 88% of the estimated sample size using the average bioequivalence analysis approach). For an IOV of 30%, and to target a power of at least 80%, 12 subjects are necessary using a cut-off of 35 and 41 (corresponding to 38% of the estimated sample size using the average bioequivalence analysis approach). For the highest tested variability (45%) and to target the same power level of at least 80%, pilot studies may be performed with 14 subjects (using a cut-off of 35) or 20 subjects (using a cut-off of 41), which correspond to 21% and 30%, respectively, of the estimated sample size using the average bioequivalence analysis approach.

Moreover, considering that none of the tested f_2 factor cut-offs inflated type I error rate (a maximum type I error of only 1% was observed for A_{mean} f_2 factor with a cut-off of 35 for simulations performed using an IOV of 45% in V [Table 6]), the authors suggest the use of a cut-off of 35 instead of 41 and 50 for the f_2 factor methodology, under the simulated conditions frame.

Despite minor differences were observed for f_2 factor derived from the A_{mean} and G_{mean} pharmacokinetic profiles, the G_{mean} f_2 factor using a cut-off of 35 was the method with the best relationship between avoiding type I and type II errors. It was also the method with higher accuracy and a better relationship between outcomes and predictions. Nevertheless, simulations are needed with more extreme scenarios (e.g., a true GMR of 90% and 80%) to better define a cut-off for this method.

A correlation between G_{mean} f_2 factor and GMR and the absolute true mean difference of *ln*-transformed test and reference C_{max} (i.e., LSM) is shown in Figure 12. The higher the absolute true mean difference of *ln*-transformed test and reference C_{max}, the lower the f_2 factor. Moreover, this figure also shows that more accurate GMR and f_2 factor estimates are obtained with the increase of the number of simulated subjects in the trial (for true bioequivalent simulations, GMR is 100% and f_2 factor is 70; for true bioinequivalent simulations, GMR is 70% and f_2 factor is 20).

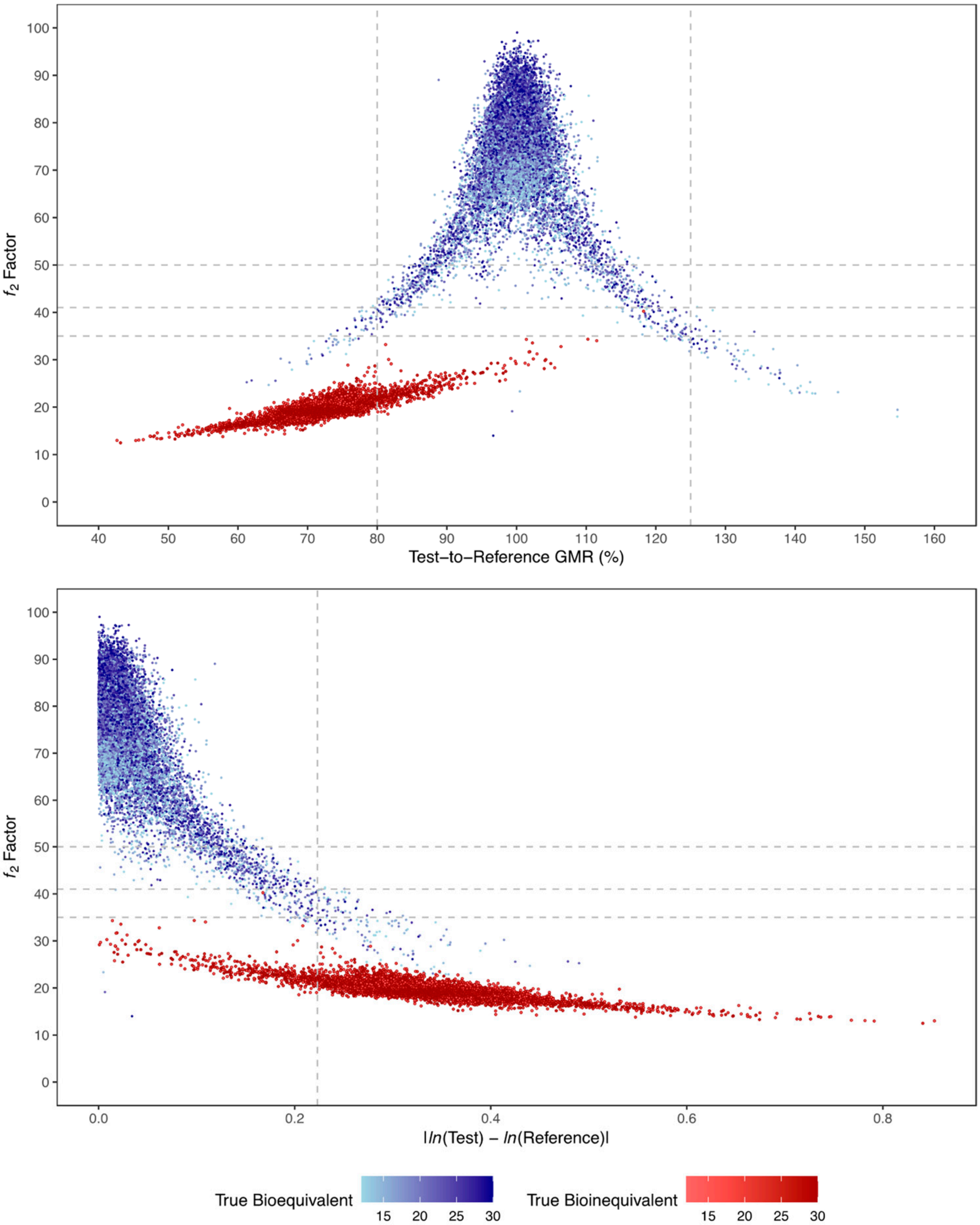

Figure 12. Relationship between G_{mean} f_2 factor and test-to-reference GMR (above) or absolute LSM difference (below), and number of subjects (colour gradient), for all simulated true bioequivalent (blue) and true bioinequivalent (red) studies. Vertical dotted lines correspond to the maximum 20% difference between test and reference formulations, tested by the average bioequivalence approach. Horizontal dotted lines correspond to the tested cut-off values for f_2 of 50, 41, and 35.

Based on the results of this work, the authors propose in Figure 13 a decision tree with a rationale for sample size and analysis approach to be followed when planning a pilot BA/BE trial for drugs characterized by a median t_{max} of approximately 2 to 4 h.

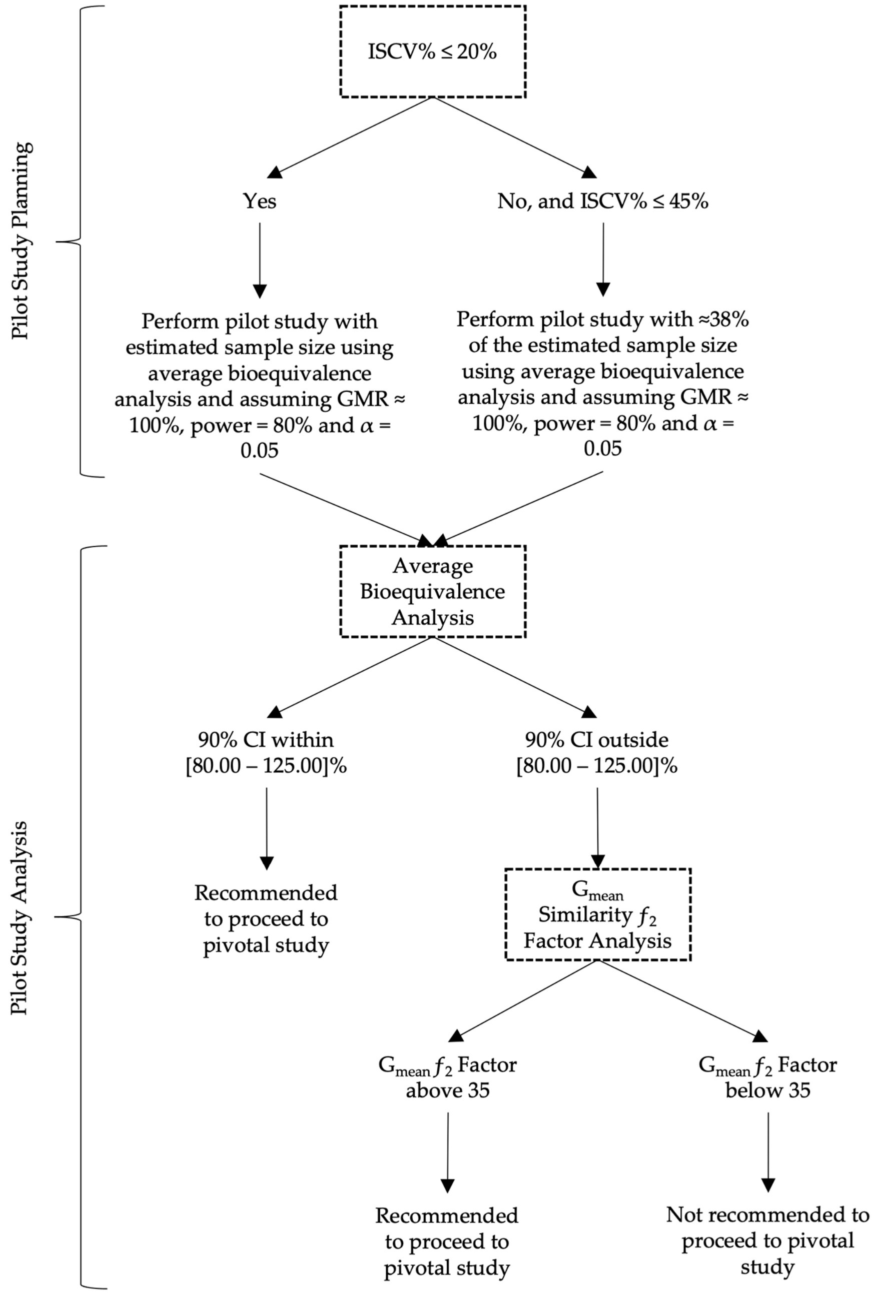
Pilot Study Planning
ISCV% ≤ 20%
Yes
No, and ISCV% ≤ 45%
Perform pilot study with estimated sample size using average bioequivalence analysis and assuming GMR ≈ 100%, power = 80% and α = 0.05
Perform pilot study with ≈38% of the estimated sample size using average bioequivalence analysis and assuming GMR ≈ 100%, power = 80% and α = 0.05
Pilot Study Analysis
Average Bioequivalence Analysis
90% CI within [80.00 – 125.00]%
90% CI outside [80.00 – 125.00]%
Recommended to proceed to pivotal study
G_mean Similarity f_2 Factor Analysis
G_mean f_2 Factor above 35
G_mean f_2 Factor below 35
Recommended to proceed to pivotal study
Not recommended to proceed to pivotal study

Figure 13. Proposed decision tree for the planning and analysis of pilot BA/BE studies.

5. Conclusions

Given the uncertainty of results derived from pilot BA/BE trials performed with drug/drug products showing a considerable variability (IOV > 20%), and consequently the uncertainty on the conclusions affecting the evaluation of the potential of a test formulation to be bioequivalent to a reference formulation on the rate of drug absorption, the authors have proposed alternative approaches to the average bioequivalence methodology that is generally applied to pilot studies to overcome and reduce the uncertainty and to help pharmaceutical companies on the decision to go forward to pivotal bioequivalence studies. The G_{mean} f_2 factor using a cut-off of 35 was found to be most appropriate method in the simulation conditions frame, enabling them to more accurately conclude on the potential of test formulations, with a reduced sample size. For simplification, a decision tree is also proposed for an appropriate planning of the sample size and subsequent analysis approach to be followed in pilot BA/BE trials.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/pharmaceutics15051430/s1, Descriptive Statistics of Simulated Pharmacokinetic Data.

Author Contributions: Conceptualization, S.C.H. and N.E.S.; methodology, S.C.H. and N.E.S.; software, S.C.H.; validation, S.C.H., J.A., P.P. and N.E.S.; formal analysis, S.C.H.; investigation, S.C.H.; data curation, S.C.H.; writing—original draft preparation, S.C.H.; writing—review and editing, J.A., P.P. and N.E.S.; visualization, S.C.H.; supervision, L.A. and N.E.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. European Medicines Agency (EMA). Guideline on the Investigation of Bioequivalence (CPMP/EWP/QWP/1401/98 Rev. 1/Corr **). London. 20 January 2010. Available online: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf (accessed on 15 September 2022).
2. U.S. Food and Drug Administration (FDA). Guidance for Industry: Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted under an ANDA. Draft Guidance. August 2021. Available online: https://www.fda.gov/media/87219/download (accessed on 15 September 2022).
3. Chow, S.C.; Liu, J. *Design and Analysis of Bioavailability and Bioequivalence Studies*; Chapman and Hall/CRC: New York, NY, USA, 2008; ISBN 9780429140365.
4. Pan, G.; Wang, Y. Average Bioequivalence Evaluation: General Methods for Pilot Trials. *J. Biopharm. Stat.* **2006**, *16*, 207–225. [CrossRef] [PubMed]
5. Fuglsang, A. Pilot and Repeat Trials as Development Tools Associated with Demonstration of Bioequivalence. *AAPS J.* **2015**, *17*, 678–683. [CrossRef] [PubMed]
6. Moreno, I.; Ochoa, D.; Román, M.; Cabaleiro, T.; Abad-Santos, F. Utility of Pilot Studies for Predicting Ratios and Intrasubject Variability in High-Variability Drugs. *Basic Clin. Pharmacol. Toxicol.* **2016**, *119*, 215–221. [CrossRef] [PubMed]
7. U.S. Food and Drug Administration (FDA). Guidance for Industry: Statistical Approaches to Establishing Bioequivalence. Draft Guidance. December 2022. Available online: https://www.fda.gov/media/163638/download (accessed on 6 March 2023).
8. European Medicines Agency (EMA). Questions & Answers: Positions on Specific Questions Addressed to the Pharmacokinetics Working Party (PKWP) (EMA/618604/2008 Rev. 13). 19 November 2015. Available online: https://www.ema.europa.eu/en/documents/scientific-guideline/questions-answers-positions-specific-questions-addressed-pharmacokinetics-working-party_en.pdf (accessed on 17 September 2022).
9. European Medicines Agency (EMA). Guideline on the Investigation of Bioequivalence—Annex I (EMA/582648/2016). 21 September 2016. Available online: https://www.ema.europa.eu/en/documents/other/31-annex-i-statistical-analysis-methods-compatible-ema-bioequivalence-guideline_en.pdf (accessed on 15 September 2022).
10. Schuirmann, D.J. A Comparison of the Two One-Sided Tests Procedure and the Power. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680. [CrossRef] [PubMed]
11. Chow, S.C. Bioavailability and Bioequivalence in Drug Development. *Wiley Interdiscip. Rev. Comput. Stat.* **2014**, *6*, 304–312. [CrossRef] [PubMed]
12. Bonate, P.L. *Pharmacokinetic-Pharmacodynamic Modeling and Simulation*, 2nd ed.; Springer: New York, NY, USA, 2011; ISBN 978-1-4419-9485-1.
13. Pigeot, I.; Hauschke, D.; Shao, J. The Bootstrap in Bioequivalence Studies. *J. Biopharm. Stat.* **2011**, *21*, 1126–1139. [CrossRef] [PubMed]
14. Labes, D.; Schütz, H.; Lang, B. PowerTOST: Power and Sample Size for (Bio)Equivalence Studies. 2010. Available online: https://github.com/Detlew/PowerTOST (accessed on 20 May 2022).
15. Pigeot, I. The Bootstrap Percentile in Food and Drug Administration Regulations for Bioequivalence Assessment. *Drug Inf. J.* **2001**, *35*, 1445–1453. [CrossRef]
16. Moore, J.W.; Flanner, H.H. Mathematical Comparison of Curves with an Emphasis on in Vitro Dissolution Profiles. *Pharm. Technol.* **1996**, *20*, 64–74.
17. Shah, V.P.; Tsong, Y.; Sathe, P.; Liu, J.-P. In Vitro Dissolution Profile Comparison—Statistics and Analysis of the Similarity Factor, F2. *Pharm. Res.* **1998**, *15*, 889–896. [CrossRef] [PubMed]

18. US Food and Drug Administration (FDA). Guidance for Industry: Waiver of In Vivo Bioavailability and Bioequivalence Studies for Immediate-Release Solid Oral Dosage Forms Based on a Biopharmaceutics Classification System. December 2017. Available online: https://www.gmp-compliance.org/files/guidemgr/UCM070246.pdf (accessed on 3 October 2022).
19. US Food and Drug Administration (FDA). Guidance for Industry: Dissolution Testing of Immediate Release Solid Oral Dosage Forms. August 1997. Available online: https://www.fda.gov/media/70936/download (accessed on 3 October 2022).
20. Kuhn, M. Building Predictive Models in R Using the Caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
21. Chicco, D.; Tötsch, N.; Jurman, G. The Matthews Correlation Coefficient (MCC) Is More Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-Class Confusion Matrix Evaluation. *BioData Min.* **2021**, *14*, 13. [CrossRef] [PubMed]

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