Pharm Res (2015) 32:135–143
DOI 10.1007/s11095-014-1450-z

RESEARCH PAPER

Inflation of the Type I Error: Investigations on Regulatory Recommendations for Bioequivalence of Highly Variable Drugs

Meinolf Wonnemann • Cornelia Frömke • Armin Koch

Received: 1 February 2014 / Accepted: 2 July 2014 / Published online: 18 July 2014

ABSTRACT

Purpose We investigated different evaluation strategies for bioequivalence trials with highly variable drugs on their resulting empirical type I error and empirical power. The classical 'unscaled' crossover design with average bioequivalence evaluation, the Add-on concept of the Japanese guideline, and the current 'scaling' approach of EMA were compared.

Methods Simulation studies were performed based on the assumption of a single dose drug administration while changing the underlying intra-individual variability.

Results Inclusion of Add-on subjects following the Japanese concept led to slight increases of the empirical α-error (≈7.5%). For the approach of EMA we noted an unexpected tremendous increase of the rejection rate at a geometric mean ratio of 1.25. Moreover, we detected error rates slightly above the pre-set limit of 5% even at the proposed 'scaled' bioequivalence limits.

Conclusions With the classical 'unscaled' approach and the Japanese guideline concept the goal of reduced subject numbers in bioequivalence trials of HVDs cannot be achieved. On the other hand, widening the acceptance range comes at the price that quite a number of products will be accepted bioequivalent that had not been accepted in the past. A two-stage design with control of the global α therefore seems the better alternative.

KEY WORDS bioequivalence · highly variable drug · pharmacokinetic · replicate design · scaling

ABBREVIATIONS

ANOVA	Analysis of variances
AUC	Area under the curve
BE	Bioequivalence
C_{max}	Maximum drug concentration
CV	Coefficient of variance
CV_{ANOVA}	Coefficient of variance calculated from the residual error in an ANOVA
EMA	European Medicines Agency
FDA	Food and Drug Administration
GMR	Geometric mean ratio
HVD	Highly variable drug
k	Regulatory constant of 0.760
L	Lower acceptance limit for bioequivalence
S_w	Residual error of an ANOVA calculation
S_{WR}	Residual error of an ANOVA calculation of two Reference administrations in a partial replicate study design
U	Upper acceptance limit for bioequivalence

INTRODUCTION

Crossover bioavailability and bioequivalence studies are performed in order to give a reliable prediction of the therapeutic effect of a dosage form. Medicinal products containing the same active substance are tested for their comparable therapeutic performance by comparing their drug concentrations mostly by measuring the concentration in the blood compartment. Rate and extent of bioavailability after administration are used to describe the *in vivo* behaviour of a substance. While the area under the concentration time curve (AUC) reflects the extent of exposure, the peak exposure, i.e. the maximum

Meinolf Wonnemann and Cornelia Frömke contributed equally to this project and should be considered co-first authors.

M. Wonnemann · C. Frömke · A. Koch
Institut für Biometrie, Medizinische Hochschule Hannover
OE 8410, 30625 Hannover, Germany

M. Wonnemann (✉)
Mainkurstrasse 18, 60385 Frankfurt a.M, Germany
e-mail: meinolf.wonnemann@gmx.de

C. Frömke
Institut für Biometrie, Epidemiologie und Informationsverarbeitung
Stiftung Tierärztliche Hochschule Hannover
Bünteweg 2, 30559 Hannover, Germany

of a curve (C_{max}) together with its respective time point of occurrence (t_{max}) better represent the rate aspect of bioavailability. AUC and C_{max} are both strongly influenced by the absorption rate and therefore used in statistical bioequivalence assessments to compare different drug formulations.

According to the current European 'Guideline on the Investigation of Bioequivalence' (1) the recommended mode of statistical evaluation is an analysis of variance (ANOVA) on logarithmic data considering pre-defined fixed and random effects. The resulting 'residual error' — the variability of interest — is assigned to the 'intraindividual variability' as studies are generally performed as a crossover trial. An assessment of average bioequivalence is then based on the point estimates and 90% confidence intervals, i.e. on the ratio of the population geometric means of the Test and the Reference product. In order to demonstrate bioequivalence, both limits of the interval of the product ratio need to be contained within the acceptance limits of −0.22314 to 0.22314 which is 0.80–1.25 or 80–125% on the original scale. This is equivalent to the two one-sided *t*-test procedures based on the alternative hypothesis of bioequivalence (2,3).

Concept of the Current European Guideline for Highly Variable Drug Products

Highly variable drugs (HVDs) are drug substances demonstrating an extraordinary high intraindividual variability in a crossover design. By definition HVDs are drugs which display an intraindividual variability of $CV_{ANOVA}\% > 30\%$. This coefficient is calculated by the following formula from the residual error (S_w) of an ANOVA evaluation:

$$CV_{ANOVA}\% = 100 \cdot \sqrt{e^{S^2_w} - 1}$$

Due to the higher intraindividual variability of their target parameters, HVDs used to have more difficulties to fulfill the authorities' requirements for acceptance of bioequivalence when administered as single doses (4). The width of a resulting confidence interval of an HVD is proportional to the intraindividual variability, and consequently more subjects need to be included in a trial to obtain tighter confidence intervals. Therefore, sample sizes of these single dose studies were high and sometimes even higher than proposed by proper sample size estimation in order to narrow the resulting confidence intervals as much as possible. The goal of only 'reasonably sized single dose studies' in HVDs could therefore not be achieved.

However, the current European guideline now offers a further strategy concerning the bioequivalence assessment of HVD products. In the case of immediate release products and single dose administration, the European guideline allows a widening of the regulatory acceptance limits for the target parameter C_{max}. The C_{max} value generally is the more variable of the two target parameters AUC and C_{max} in a bioequivalence assessment. Drug formulations of highly variable substances may be investigated in a replicate study design in case '*a wider difference in C_{max} is considered clinically irrelevant based on a sound clinical justification*' (1). The replicate design may imply a crossover double administration of the Reference product beneath the Test administration. ANOVA analysis of the two Reference administrations and the residual variability obtained then serve for the calculation of the intraindividual variability of the two References. The Reference variability serves in a final step for calculation of acceptance limits for the following comparison against the Test product in a data driven way.

The so called 'scaling' procedure follows the formula

$$[L, U] = \exp[\pm k \cdot S_{WR}],$$

where L and U are the lower and upper acceptance limits for bioequivalence, k is a regulatory constant of 0.760 and S_{WR} represents the intraindividual variability determined in the comparison of the two Reference administrations. Following the guideline, all drug substances below an intraindividual variability of CV_{ANOVA} 30% remain at the equivalence bounds of 80% to 125% while all between 30% and 50% may follow the 'scaled' approach. Therefore, according to this concept, the higher the Reference variability and the resulting coefficient of variation (CV_{ANOVA} %) will be, the wider the bioequivalence limits for the comparison of Test *vs.* Reference will be set. However, the maximum allowed widening of acceptance limits is 69.84% to 143.19% which is equal to a variability of 50%. Additionally, a point estimate between the 80% and 125% limits is required to fulfill the definitions of bioequivalence. Resulting 'scaled' lower and upper limits are presented in Table I.

Several biometrical publications have investigated the idea of the 'scaled' average bioequivalence approach. Aspects dealt with include the comparison of this approach to the ordinary 'unscaled' process with limits of 80%–125% (5–8,10) or with

Table I Bioequivalence Limits of the 'Scaled' Average Approach (1)

Intraindividual CV [%]	Lower limit [%]	Upper limit [%]
30	80.00	125.00
35	77.23	129.48
40	74.62	134.02
45	72.15	138.59
50	69.84	143.19

$$CV = \sqrt{e^{S^2_{WR}} - 1}$$

differences of average bioequivalence and individual equivalence calculations (9–11). Moreover, the robustness of this 'scaled' concept regarding possible outliers which could influence the estimations of variability was investigated (5). The issue of the relation of different point estimates (6), intra-individual variability and the empirical statistical power in different and also replicate study designs was intensively investigated (12–15). Finally, the concepts of EMA and the FDA, to establish a 'scaling' strategy were compared in several publications (6,8,16,17). Good overall insight on the current European and US American regulatory strategies discussed together with simulation experiments was given by Endrenyi and Tothfalusi (18) and Davit (19). The authors moreover discuss the switching limit of CV_{ANOVA} 30% and the additional requirement of a point estimate inside the limit of 80%–125%.

Tothfalusi *et al.* (5) presented exemplary data of the empirical statistical power determined in simulation studies assuming a CV_{ANOVA} of 40%. When sample sizes of only N = 24 or N = 12 subjects were used, the empirical power rate was approximately 87–88% with a geometric mean ratio (GMR) of 1.00. At a GMR of 1.05, it was still about 84 to 85%, reflecting a strong robustness in terms of the producer risk of false negative results.

Concept of the Japanese Guideline for Highly Variable Drug Products

An Add-on strategy is the method of choice described in the current Japanese Guideline for Bioequivalence Studies of Generic Products (20). '*An add-on subject study can be performed using not less than half the number of subjects in the initial study*', which can be used in the case the bioequivalence assessment in an HVD failed. It is further stated that a sample size of 20 subjects for the initial part plus an Add-on of 10 subjects may suffice if investigated products are similar in dissolution rates and average AUC and C_{max} values.

It is the aim of this work to compare the two HVD bioequivalence approaches of Europe and Japan to the classical 'unscaled' evaluation in terms of the major biometrical parameters while changing intraindividual variability and sample sizes. The focus is to estimate the resulting consumer and producer risks when studies are designed and calculated as proposed in these guidelines and to elaborate on the impact of decision making about bioequivalence of generic drugs.

MATERIALS AND METHODS

Simulation studies were conducted based on the assumption of a single dose drug administration and log-normally distributed data with different geometric mean Test/Reference ratios. At first, we investigated the resulting empirical type-I-error, i.e. the α-error rate at the geometric mean ratio (GMR) of 1.25, the border of the acceptance range for the classical 'unscaled' average bioequivalence approach by calculating the rejection rate in the simulation studies. Afterwards, we investigated the error rate under the assumption of the Japanese Add-on strategy. Finally, the empirical rejection rates at a GMR of 1.25 and at the respective 'scaled' limits of the European concepts were calculated. With 'scaling' the rejection rate at a GMR of 1.25 cannot be called a type I error, as with scaled limits applied it is not a rejection of the null - hypothesis. However, the empirical rate allows for a direct comparison with the strategy of an unscaled bioequivalence assessment.

We investigated the impact of changing the intraindividual variability and sample sizes following all three concepts.

In a second step we investigated empirical power of the classical 'unscaled' approach, i.e. the empirical rejection rate of H_0. Subsequently, we investigated the proposal of Japan and the influence of increasing sample sizes in the Add-on group on empirical power. Empirical power was finally also investigated following EMA's 'scaled' evaluation approach. All three approaches were again investigated while changing intra-individual variability and sample size.

Sample sizes for the classical 'unscaled' approach, the Japanese Add-on and the 'scaled' EMA strategy were estimated considering intraindividual variabilities in the range of CV_{ANOVA} 10% to 60%, an α-error rate of 5%, a minimum power of 80%, and the respective GMR ratio on the logarithmic scale. Sample sizes used were in accordance with the literature (21,22).

For determination of the empirical rejection rates, the developed macros were run considering values of intraindividual variability in the range of CV_{ANOVA} = 20 to 50% and a GMR in the range of 1.00 to 1.40 for the 'unscaled' and the Add-on design on the logarithmic scale. For the 'scaled' approach of EMA, we chose a wider range of CV_{ANOVA} = 20 to 55% and a GMR of 1.00 to 1.60 in order to investigate also the development beyond the regulatory upper cap of the scaling approach of CV_{ANOVA} 50%. Sample sizes were adapted accordingly.

In all simulation studies we simulated normally distributed random numbers (i.e. we simulated data after logarithmic transformation, as usually kinetic parameters are lognormally distributed). Data simulations and corresponding evaluations were performed disregarding a pre-set interindividual variability. Moreover, variability of a subject-by-formulation-interaction was not considered and set to CV = 0%. Macros for data creation did not consider a possible data relation/data correlation when belonging to one subject under different conditions; the correlation, i.e. the covariance structure, was set to 0. When investigating both the empirical power 5000 simulations, and when investigating the empirical type I error

rate, 10000 simulations were performed for each condition and data point respectively.

When simulating the Add-on strategy according to the Japanese guideline, at first a regular AB|BA crossover design was assumed for the initial part of the studies. All simulated studies passing the acceptance limits were counted and excluded from further simulation procedures. In all studies where limits were exceeded, additional Add-on subjects ($N_{total}=N+N/2$ subjects) were included in the design. Finally, the resulting rates were summarized to a final empirical rate.

SAS Macros for operation of simulation studies were valid for analyses, i.e. the maximum and mean simulation error when investigating the type I error rate had been + 0.40% and + 0.16%, respectively.

Macros for operation of simulation studies were created in SAS® software (version 9.2) (23). Graphs were prepared with GraphPad Prism Software (version 5.04) (24).

RESULTS

Determination of the Empirical Type I Error Rate

At first we investigated the empirical α-error rate with the simple 'unscaled' average bioequivalence evaluation of an AB|BA design at the classical limit of GMR 1.25 while increasing sample sizes in the variability range of CV_{ANOVA} between 10% and 55%. As shown in Fig. 1 the test is conservative in the beginning and finally, all simulated curves converged at an empirical α - level close to 5% as expected. The lower the variability, the earlier the convergence to the pre-set α-level of 5% was achieved.

In the case of the Japanese design, the empirical α-error rate of the first step was similar to the classical crossover design i.e. all values finally converged at a level of 5% (data not shown) when investigated at the GMR of 1.25. However, when Add-on subjects were included in the magnitude of half of the subjects of the first step ($N_{total}=N+N/2$ subjects), the α-error increased. Curves of the simulation studies all converged at about 7.5% and, again, the lower the intraindividual variabilities had been, the earlier the error level was reached with lower total sample sizes (see Fig. 2). Therefore, independent of the underlying intraindividual variability and sample size, an increase of the type I error rate after convergence of about + 2.5% above the preset limit of α=5% is to be anticipated in the investigated range.

Finally, we investigated the empirical rejection rate for the European 'scaled' proposal of EMA. First, we investigated the empirical rate at the classical bioequivalence border of GMR 1.25 (see Fig. 3) With scaled limits applied this rate cannot be called a type I error rate as at the GMR of 1.25 it is still not a rejection of the null - hypothesis. However, the empirical rate allows for a direct comparison with the strategy of an unscaled bioequivalence assessment. Increases in the rate thus directly reflect the increased probability to accept equivalence of two products under the respective configuration of CV_{ANOVA} and sample size. At the lowest variability of CV_{ANOVA} of 25% the results were as expected, i.e. the error rate converged at a level of 5% as the 'scaling' is still not applied. At the cut-off point of CV_{ANOVA} of 30%, the rate was slightly elevated to 7.05%. When investigating intraindividual variabilities in the range of CV_{ANOVA} of 30% to 50% where 'scaled' bioequivalence limits have to be applied, a tremendous increase of the empirical rejection rate with increasing sample sizes could be noted. Moreover, contrasting with the other designs, this rejection rate almost did not converge to a certain error level in the investigated sample size range. With a total sample size of 24 subjects, the rejection rate was approximately 24.22% with a variability of CV_{ANOVA} of 50%. A further increase in variability to CV_{ANOVA} of 55% and 60% did not cause a further elevation, due to the upper pre-set limit of the scaling procedure.

To obtain further insight into the pattern of α-error development while using appropriate sample sizes, we again simulated the error rate at a GMR of 1.25 for each design while changing the intraindividual variability in small steps. Results are displayed in Fig. 4. Calculations confirmed an increase of

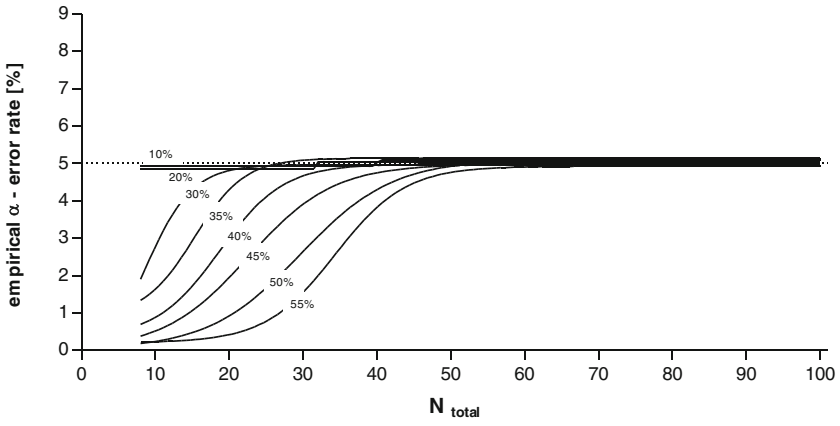

Fig. 1 Empirical α-error rate at a GMR of 1.25 with increasing sample size and intraindividual variability—rate of H_0 rejections after 10000 simulations—AB|BA crossover design with 'unscaled' average bioequivalence evaluation.

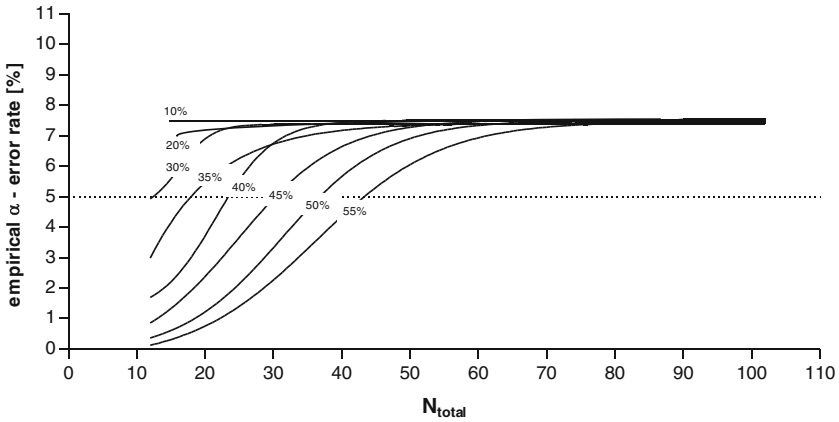

Fig. 2 Empirical α-error rate at a GMR of 1.25 with increasing sample size and intraindividual variability—rate of H_0 rejections after 10000 simulations—Add-on design according to the Japanese guideline (Add-on = +N/2 subjects) with 'unscaled' average bioequivalence evaluation.

the rate to about 7.5% with the Japanese Add-on strategy. Sample size estimation for the replicate EMA design was in accordance with that of Tothfalusi *et al.* (22). The rejection rate was 5.12% at a CV_{ANOVA} of 20% followed by a slight further increase. Starting with CV_{ANOVA} of 30% and the 'scaled' evaluation, the rejection rate rapidly increased with increasing variability to about 12, 19, 24, and 27% at CV_{ANOVA} values of 35, 40, 45, and 50%.

To complete and confirm the latter results, finally, we performed simulations at a GMR of 1.25 and at the respective 'scaled' bioequivalence limits pre-set by the guideline. Intraindividual variabilities and samples sizes given by Tothfalusi *et al.* (22) were considered. Results are displayed in Table II. Simulations confirmed the dramatic increase of the rejection rate at a GMR of 1.25. The empirical rejection rate rose from 7.05% at CV_{ANOVA} of 30% to 26.56% at 50%. Surprisingly, when 'scaled' bioequivalence limits were set as bioequivalence limits, the highest rejection rate observed was at the lowest variability investigated. With CV_{ANOVA} of 30%, it was 7.05%, and it was still 5.39% with a variability of 40%. Therefore, taking into account a simulation error of roughly 0.5%, and the fact that our simulations are based on uncorrelated data and do not consider interindividual variabilities, one may doubt that an α-error of 5% is controlled even with the pre-set 'scaled' limits, at least for variabilities close to the cut-off point of CV_{ANOVA} of 30%.

Determination of the Empirical Power

Power results of the simple AB|BA two—period crossover design at a GMR of 1.00 were in the range of 84.04% (at a variability of CV_{ANOVA} of 20%) to 78.08% (at a variability of 50%). At a GMR of 1.10 the rate was reduced to 53.12%–49.96% in the variability range of 20% to 50% (see Fig. 5).

In the case of the Japanese design, the empirical power increased to 97.00% at a variability of CV_{ANOVA} of 20% and was still 95.65% with a variability of 50% at a GMR of 1.00 when Add-on subjects were included. At a GMR of 1.10 the rate was 73.85% with CV_{ANOVA} of 20% and 69.55% at 50% (see Fig. 1).

Inclusion of more subjects in the Add-on part resulted in an additional enormous increase in empirical power rates. With a CV_{ANOVA} of 40% and a GMR of 1.00 the power could be increased from 82.80% up to 99.50% with N = 56 in the Add-on part (for details please refer to (22)).

Finally, we investigated the empirical H_0 rejection rate for the European 'scaled' design of EMA. At a GMR of 1.00 we determined 81.66% for variability CV_{ANOVA} of 30% and

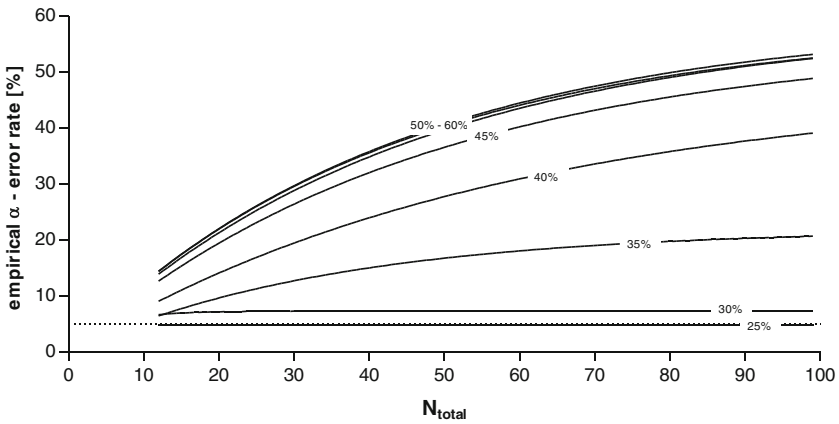

Fig. 3 Empirical rejection rate at a GMR of 1.25 with increasing sample size and intraindividual variability—rate of rejections after 10000 simulations—replicate three period study designs according to EMA with 'scaled' average bioequivalence evaluation.

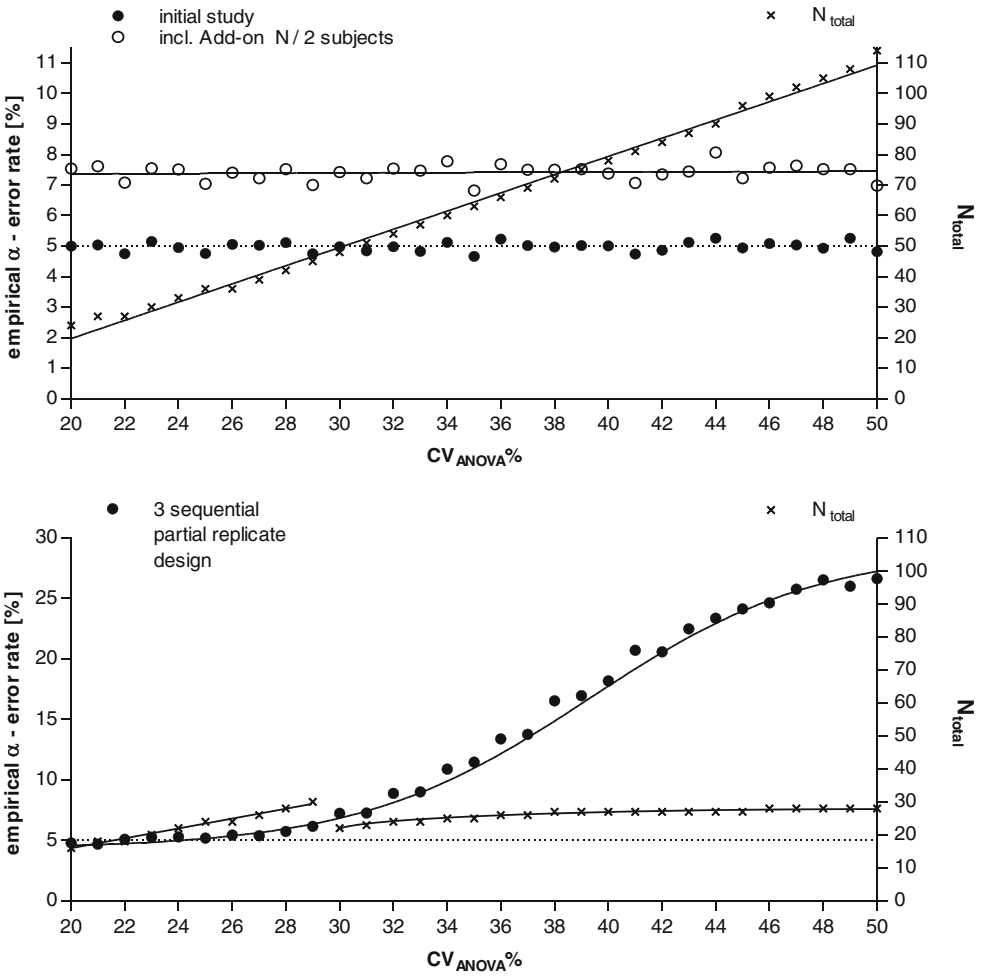

Fig. 4 Empirical α-error rate at a GMR of 1.25 *vs.* CV_{ANOVA} at affiliating sample size—rejection rate after 10000 simulations—Japanese design (*top*), and replicate 3 period design with 'scaled' evaluation according to EMA (*bottom*).

76.12% in the case of 55%. Increasing the product difference to a GMR of 1.10 resulted in empirical power values of 53.88% at 30% and 67.42% at a variability of 50% (see Fig. 1).

Table II Rejection Rate After 10000 Simulations at a Fixed GMR of 1.25% (Upper Part) and Empirical α-Error Rate at the 'Scaled' BE Limits (Lower Part) According to EMA with Increasing Intraindividual Variability

CV_{ANOVA} [%]	*N*	GMR	Empirical rejection rate [%]
30	22	1.250	7.05
35	25	1.250	11.82
40	27	1.250	18.74
45	27	1.250	23.57
50	28	1.250	26.56
CV_{ANOVA} [%]	*N*	GMR	Empirical α-error rate [%]
30	22	1.250	7.05
35	25	1.295	5.58
40	27	1.340	5.39
45	27	1.386	4.25
50	28	1.432	3.51

DISCUSSION

The results of our simulation studies in HVDs demonstrate substantial differences in empirical power and α-error rates between the classical 'unscaled' AB|BA crossover design, the concept of the Japanese approach (20), and the new 'scaling' strategy of the current European guideline (1).

Simulation of the empirical power following these strategies showed a marked increase in power in comparison to the classical AB|BA design where Add-on subjects were included according to the Japanese approach, while it was independent of underlying variability. When sample sizes in the Add-on part were further escalated, power values could be increased even more (25). Empirical power rates of the partial replicate EMA design with 'scaled' evaluation showed dependency on the underlying variability, but, in comparison to the classical approach, they were more stable and markedly higher in a GMR range of 1.00 to 1.10. When product difference was rising, empirical power notably declined in the classical approach, while it was less influenced with the EMA approach. Therefore, we could mainly confirm the data of

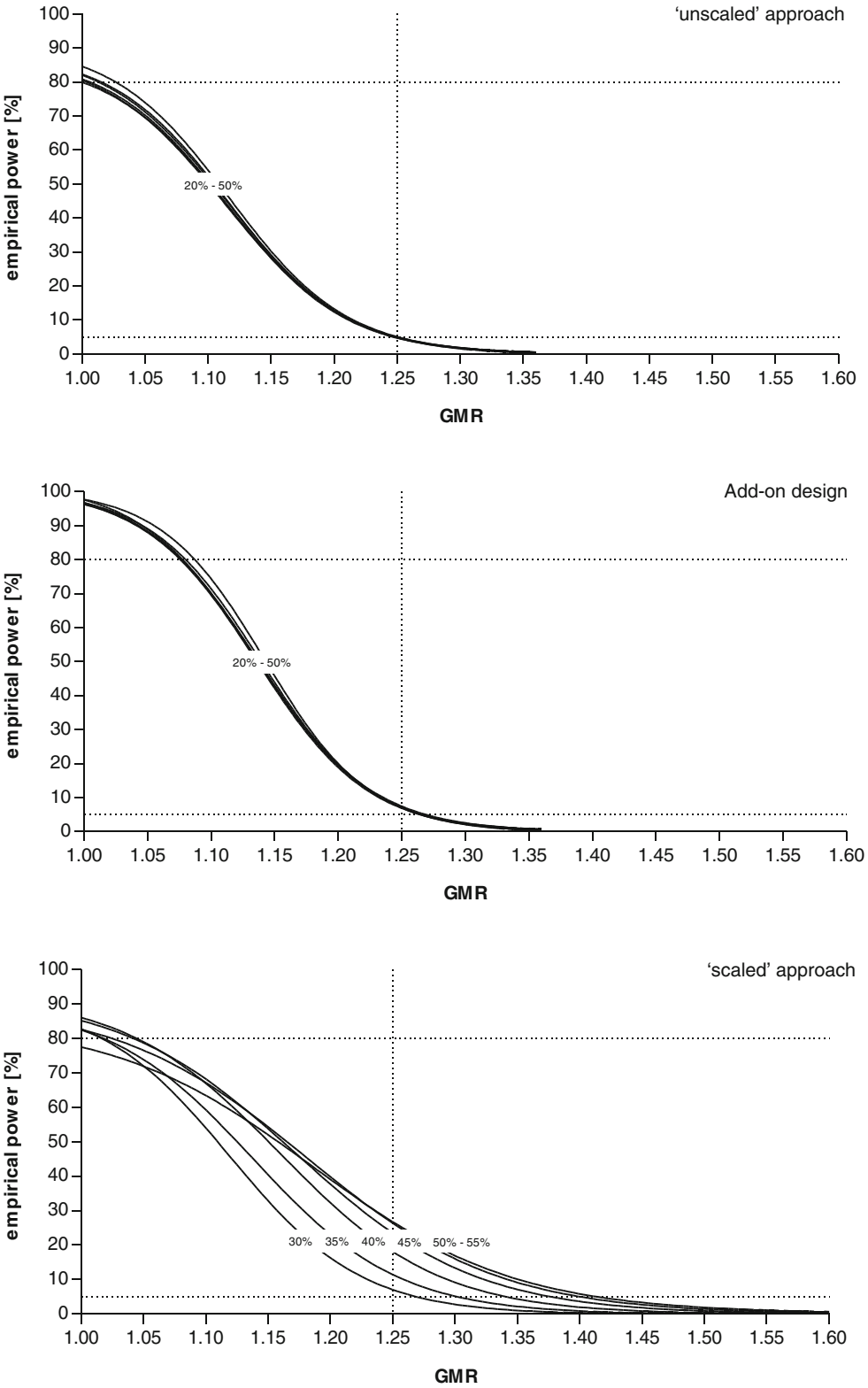

Fig. 5 Empirical power rate with increasing CV_{ANOVA} in the GMR range 1.00 to 1.60—rejection rate with the classical 'unscaled' average bioequivalence evaluation (*top*), according to the Add-on study design following the Japanese guideline (Add-on = subjects of step 1 + N/2 subjects) (*middle*), and according to the three period replicate design following the European guideline with 'scaled' average bioequivalence evaluation (*bottom*).

Tothfalusi *et al.* (5,14) and others on empirical power rates for a 'scaled' evaluation of a partial replicate study design.

We simulated the type – I error rate at a GMR of 1.25 with intraindividual variabilities in the range of CV_{ANOVA} of 10% to 60%. The results demonstrated pronounced differences between the three concepts, which can be plausibly explained by the fact that a GMR of 1.25 is under the alternative hypothesis in the 'scaled' average bioequivalence approach,

whereas it is under the null-hypothesis of the other tests. While empirical error rates were controlled at a level of 5% with all variabilities by the classical 'unscaled' evaluation method, an inclusion of Add-on subjects lead to an increase of the rates. It was stable at around 7.5% and independent of underlying intraindividual variability following the Japanese concept. Surprisingly, when investigating the rejection rate following the 'scaled' approach of EMA, we noted a tremendous increase at a GMR of 1.25 – the former border of bioequivalence. With a CV_{ANOVA} of 40% an almost fourfold with a variability of 50%, a more than fivefold increase of the rejection rate was observed. Moreover, when 'scaled' bioequivalence limits were applied, we still detected rejection rates slightly above the pre-set limit of 5%. It has to be mentioned that we detected an α-elevation up to 7.05% with a CV_{ANOVA} of 30%, which means a difference less than 1% in comparison to the error rate following the Japanese approach.

To our knowledge, this is the first time that the different proposals for demonstrating bioequivalence for HVDs have been compared at a GMR of 1.25. Comparable error rates can be found either following the approach of EMA or recommendations of the FDA which partially use a different cut-off point of CV_{ANOVA} of 25% for starting the scaling procedure (5,6,8,10,13–17).

As shown by our results, the goal of reasonable sample sizes obviously cannot be achieved with the approach of the Japanese guideline. Due to the Add-on part sample size will be higher than that of the classical average bioequivalence approach. However, because of the opportunity of a second chance to meet bioequivalence criteria, the Add-on concept reduces the risk of misjudgement and failure of trials for the price of a slightly elevated consumer risk. On the other hand, the 'scaled' procedure clearly demonstrates that a concept with stable power connected with lower reasonable sample sizes seems to be feasible in bioequivalence trials.

Yet, these advantages on the more economical side go along with a very pronounced increase of the α-error when empirical rates are compared at the GMR of 1.25. Control of sample size and the producer risk therefore leads to a more liberal position regarding a possible type I error.

In order to solve this dilemma, health agencies around the world have different ideas of how to proceed and how to conduct studies in HVDs. For a strict control of the global α-error rate, the authority of New Zealand, for example, allows a sequential design in studies with HVDs, where the sample size of the second step is based on the results of the first study step (26). However, the authority also limits the subject number to 40. To prove bioequivalence, the latter might result in more trials than necessary (assuming that all generic products submitted to regulatory authorities are bioequivalent), as the chance for demonstrating bioequivalence is markedly reduced by a restricted sample size for HVDs. More trials than necessary, however, also stand against the general effort to reduce subject numbers in bioequivalence trials.

Besides the opportunity to use the 'scaling' procedure for studies in HVDs, EMA also accepts a two-stage approach for demonstrating bioequivalence with control of the global α-level independent of the underlying intraindividual variability (1). Thus the guideline at the same time also allows conducting trials that will control the type-I-error in a strict sense.

Moreover, there are several other ideas that present a good way out of the dilemma mentioned above. One good option is the additional inclusion of biopharmaceutical data to support an assessment of bioequivalence. The Japanese guideline proposes an Add-on of not less than N/2 subjects of the first study step. However, bioequivalence will be also accepted even though this strategy did not meet the acceptance limits, if:

> "*1) the total sample size of the initial bioequivalence study is not less than 20 (n = 10/group) or pooled sample size of the initial and add-on subject studies is not less than 30, 2) the differences in average values of logarithmic parameters to be assessed between two products are between log(0.9) and log(1.11), and 3) dissolution rates of test product are equivalent to those of the reference product.*"

Here, acceptance criteria comprise the usual acceptance limits, a constraint on the point estimate and additional biopharmaceutical dissolution data to support a presumption of equivalence.

CONCLUSIONS

In conclusion, the 'unscaled' average bioequivalence evaluation with an underlying two period AB|BA design as well as the concept of the Japanese guideline indeed does not contribute to the aim of reducing subject numbers in bioequivalence trials of HVDs. However, the 'scaling' strategy of the current European guideline with lower sample sizes contains the risk that a relevant amount of HVDs will enter the market that would have been assessed as non-bioequivalent under the classical evaluation method. Every effort shall be appreciated finding a reasonable and practicable solution, which is not solely at the expenses of the consumer due to the mode of statistical evaluation, because in the end any widening of the acceptance range (irrespective of whether this is done with an increase in the type-I-error, or with scaling) can only be justified by firm knowledge that in the presence of higher variability larger product differences are of no clinical relevance. This knowledge, however, is rarely available.

REFERENCES

1. Guideline on the Investigation of Bioequivalence (CPMP/EWP/QWP/1401/98 Rev. 1/Corr, January 2010.
2. Hauck WW, Hauschke D, Diletti E, Bois FY, Steinijans VW, Anderson S. Choice of student's t- or Wilcoxon-based confidence intervals for assessment of average bioequivalence. J Biopharm Stat. 1997;7(1):179–89.
3. Schuirmann DJ. A comparison of the two one sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987;15:657–80.
4. Note for Guidance on the Investigation of Bioavailability and Bioequivalence (CPMP/EWP/QWP/1401/98). January 2002.
5. Tothfalusi L, Endrenyi L, Arieta AG. Evaluation of bioequivalence for highly variable drugs with scaled average bioequivalence. Clin Pharmacokinet. 2009;48(11):725–43.
6. Karalis V, Symilides M, Macheras P. On the leveling-off properties of the new bioequivalence limits for highly variable drugs of the EMA guideline. Eur J Pharm Sci. 2011;44(4):497–505.
7. Patterson SD, Zariffa NMD, Montague TH, Howland K. Non-traditional study designs to demonstrate average bioequivalence for highly variable drug products. Eur J Clin Pharmacol. 2001;57:663–70.
8. Karalis V, Sylmillides M, Macheras P. Bioequivalence of highly variable drugs: a comparison of the newly proposed regulatory approaches by FDA and EMA. Pharm Res. 2012;29(4):1066–77.
9. Howe WG. Approximate confidence limits on the mean of X + Y where X and Y are two tabled independent variables. J Am Stat Assoc. 1974;69:789–94.
10. Patnaik RN, Lesko LJ, Chen ML, Williams RL. Individual bioequivalence: new concepts in the statistical assessment of bioequivalence metrics. Clin Pharmacokinet. 1997;33:1–6.
11. Chen ML, Lesko LJ. Individual bioequivalence revisited. Clin Pharmacokinet. 2001;40(10):701–6.
12. Phillips KF. Power of the two one-sided tests procedure in bioequivalence. J Pharmacokinet Biopharm. 1990;18:137–44.
13. Tothfalusi L, Endrenyi L, Midha KK. Scaling or wider bioequivalence limits for highly variable drugs and for the special case of C(max). Int J Clin Pharmacol Ther. 2003;41(5):217–25.
14. Tothfalusi L, Endrenyi L. Limits for the scaled average bioequivalence of highly variable drugs and drug products. Pharm Res. 2003;20(3):382–9.
15. Tothfalusi L, Endrenyi L, Midha KK, Rawson MJ, Hubbard JW. Evaluation of the bioequivalence of highly-variable drugs and drug products. Pharm Res. 2002;18(6):728–33. Comment in Pharm Res 19(3):227–8.
16. Haider SH, Davit B, Chen ML, Connor D, Lee LM, Li QH, *et al.* Bioequivalence approaches for highly variable drugs and drug products. Pharm Res. 2008;25(1):237–41.
17. Haider SH, Makhlouf F, Schuirmann DJ, Hyslop T, Davit B, Conner D, *et al.* Evaluation of a scaling approach for the bioequivalence of highly variable drugs. AAPS. 2008;10(3):450–4.
18. Endrenyi L, Tothfalusi L. Regulatory conditions for the determination of bioequivalence of highly variable drugs. J Pharm Sci. 2009;12(1):138–49.
19. Davit BM, Chen ML, Conner DP, Haidar SH, Kim S, Lee CH, *et al.* Implementation of a reference-scaled average bioequivalence approach for highly variable generic drug products by the US food and drug administration. AAPS J. 2012;14(4):915–24.
20. Guideline for Bioequivalence Studies of Generic Products, 医薬品の生物学的同等性試 験ガイドライン, 発医薬品の生物学的同 等性試験ガイドライン, 発医薬品の生物学的 同等性試験ガイドライン, Ministry of Health, Labour and Welfare (MHLW). Japan; 2012.
21. Chow SC, Liu JP. Design and analysis of clinical trials. 2nd ed. Hoboken: Wiley-Interscience; 2004. p. 483. ISBN 0-471-24985-8.
22. Tothfalusi L, Endrenyi L. Sample sizes for designing bioequivalence studies for highly variable drugs. J Pharm Pharm Sci. 2012;15(1):73–84. www.cspsCanada.org.
23. SAS Institute Inc. SAS/STAT 9.2 SAS User's Guide. Cary, North Carolina, USA: SAS; 2002–2008.
24. GraphPad Prism version 5.04 for Windows. La Jolla California USA: GraphPad Software Inc.
25. Wonnemann M. Different approaches for the assessment of bioequivalence of highly variable drug products—comparison of rules given in the current European, Canadian and Japanese guidelines. Master thesis, Ruprecht-Karls University of Heidelberg; 2012.
26. Guidance notes for applicants for consent to distribute new and changed medicines and related products, New Zealand Medicines and medical devices safety authority (MedSafe) New Zealand Regulatory Guidelines for Medicines. New Zealand; 2001.