Novel Scaled Average Bioequivalence Limits Based on GMR and Variability Considerations

Vangelis Karalis,[1] Mira Symillides,[1,2] and Panos Macheras[1]

Received March 4, 2004; accepted June 10, 2004

Purpose. i) To develop novel approaches for the construction of bioequivalence (BE) limits incorporating both the intrasubject variability and the geometric mean ratio (GMR), and ii) to assess the performance of the novel approaches in comparison to several scaled BE procedures and the classic unscaled average BE.

Methods. Plots of the BE limits or the extreme GMR values accepted as a function of the coefficient of variation (CV) were constructed for published and the developed scaled procedures. Two-period crossover BE investigations with 12, 24, or 36 subjects were simulated with assumptions of a CV 10%, 20%, 30%, or 40%. The decline in the percentage of accepted studies was recorded as the true GMR for the two formulations was raised from 1.00 to 1.50. Acceptance of BE was evaluated by published and the developed scaled procedures, and, for comparison, by the unscaled average BE.

Results. Two GMR-dependent BE limits are proposed for the evaluation of average BE: i) BELscG1 with Ln(Upper, Lower BE limit) = ±[(5 − 4GMR)0.496s + Ln(1.25)], and ii) BELscG2 with Ln(Upper, Lower BE limit) = ±[(3 - 2GMR)(0.496s + Ln(1.25))], where s is the square root of the intrasubject variance. The range of BE limits becomes narrower as GMR values deviate from unity, and increases with variability. The two new approaches exhibit the highest statistical power at low CV values. At high levels of variability, BELscG1 and BELscG2 show high statistical power, as well as the lowest percentages of acceptance among the scaled methods when GMR = 1.25. The latter becomes more obvious when a large number of subjects is incorporated in the studies.

Conclusions. The GMR and CV estimates of the BE study can be used in conjunction with the GMR vs. CV plot for the assessment of average BE. The new approaches, BELscG1 and BELscG2, appear to be highly effective at all levels of variation investigated.

KEY WORDS: bioequivalence; GMR; highly variable drugs; regulatory criteria; scaled bioequivalence limits.

INTRODUCTION

The classic methodology (1,2) for the determination of bioequivalence (BE) is based on the two-period crossover design where two drug products are considered bioequivalent if the 90% confidence interval (CI) for their mean relative bioavailability (average bioequivalence, ABE) lies between the predefined limits 0.80–1.25. However, the problems of establishing BE for highly variable (HV) drugs with these constant values for BE limits, that is, with the present unscaled ABE, are well-known (3–5). For example, it is very difficult to prove BE when the intrasubject variability is high [coefficient of variation (CV) >30%], and a large number of subjects is required to achieve adequate statistical power. On the other hand, it has been realized that the unscaled ABE allows large differences between the means for drug products with low residual variability. This constitutes a potential problem of switchability for multisource formulations, each declared bioequivalent to the same reference product (6,7). A problem may also arise in the case of toxic drugs with low variability and narrow therapeutic range (7).

To overcome these difficulties, several approaches have been proposed. In order to reduce intrasubject variability, multiple dose steady-state studies have been considered (8). Replicate designs for single-dose studies, reducing the total number of subjects required, have also been proposed (2,5,8). Nevertheless, these methods increase the duration of exposure of the volunteers and moreover, potential practical problems may arise (e.g., increased incidence of subject withdrawals).

An alternative method was discussed, especially for pharmacokinetic parameters showing increased variation as peak plasma drug concentration, C_{max}, that is, widening the bioequivalence acceptance limits to prefixed constant values (0.70–1.43) (1,4,9). Additionally, a method for expanding the limits for HV drugs, based on the estimate of the intrasubject CV, was proposed; therefore, BE limits are scaled according to a fixed multiple of CV (10). Different rationales have been developed concerning the choice of the proportionality factor for scaled BE limits termed hereafter BELsc (7,10,11). Moreover, several other procedures like Individual Bioequivalence (12–16) or scaled Average Bioequivalence (ABEsc) (17,18) have been considered. It is worth mentioning, that the model for ABEsc can be readily converted to that of the BELsc. Indeed, when investigated, the two approaches yielded very similar results (18). Nevertheless, as mentioned earlier (18), the BELsc procedures could be preferred because the relevant confidence limits can be easily computed by the usual t-statistics.

In recent papers (9,11), two interesting variants of the scaled procedure were investigated. The first approach proposed a mixed method (11), using the classic unscaled ABE when drugs do not exhibit high variability, and scaled ABE for HV drugs, that is, when a preset magnitude of the variability is exceeded. The "switching" variability, σ_0, for the scaled ABE was set to 0.20, and corresponds to a proportionality constant, (Eq.10 of Ref 11), $k = \mathrm{Ln}(1.25)/\sigma_0 = 1.116$. The mixed model (11) for scaled ABE can be converted to a mixed approach of scaled BE limits, using the classic unscaled criterion up to CV 20% and scaled BE limits with a proportionality factor of 1.116, for CV over 20%. The second ap-

[1] Laboratory of Biopharmaceutics-Pharmacokinetics, School of Pharmacy, University of Athens, Athens, Greece.

[2] To whom correspondence should be addressed. (e-mail: simillidou@pharm.uoa.gr)

ABBREVIATIONS: ABE, average bioequivalence; ABEsc, scaled average bioequivalence; AUC, area under the concentration-time curve; BE, bioequivalence; BELsc, scaled bioequivalence limit; C_{max}, maximum observed plasma concentration; CI, confidence interval; CV, coefficient of variation; Diff, difference of the logarithmic values of test and reference formulation; $Diff_{max}$, the maximum value for Diff to declare bioequivalence; GMR, geometric mean ratio; GMR_{max}, the maximum accepted GMR value for two products to be bioequivalent; GMR_{min}, the minimum accepted GMR value for two products to be bioequivalent; HV, highly variable; MSE, mean square error of analysis of variance; N, number of subjects participating in bioequivalence trials; R, reference formulation; s, intrasubject variability; σ_0, switching variability; T, test formulation.

0724-8741/04/1000-1933/0 © 2004 Springer Science+Business Media, Inc.

proach is dealing with the deviations of geometric mean ratios observed while using scaled procedures. Because large differences between the means can be accepted by scaled methods with substantial probabilities, an additional regulatory criterion was imposed concomitantly with the classic 90%CI in BE limits (9). This secondary criterion suggests that the estimated ratio of geometric means (GMR) should be constrained in the range 0.80–1.25.

In this study, we present a new approach for the construction of scaled BE limits criteria. In order to overcome the drawbacks of the classic and the scaled BE limits already appearing in literature, we propose a different rationale for the development of the scaled limits. Our approach is based on the incorporation of a GMR constraint criterion for the construction of BE limits. In this context, both the GMR and residual variability estimates are used for the formulation of acceptance limits. The performance of the resulting new scaled procedures is evaluated and compared with the performance of the classic unscaled ABE and several scaled methods proposed in the literature.

METHODS

The usual procedure for determining the average bioequivalence of two formulations implies that the means of a logarithmically transformed metric, (such as LnC_{max} or LnAUC) for the test and the reference formulations are contrasted. Bioequivalence is declared if the 90% confidence interval (CI) for the difference of log means is within preset bioequivalence limits. Assuming two-treatment, two-period, crossover BE studies, with equal numbers of subjects in each sequence, the upper and lower limits of the 90%CI are calculated according to Eq. 1:

$$\begin{aligned}&\text{Upper, Lower limits of the 90\% CI}\\&\quad= \exp(\text{Diff} \pm t_{0.05,N-2}\sqrt{s^2 2/N})\end{aligned} \tag{1}$$

where Diff is the difference of test and reference means of the metric m_T and m_R, respectively, that is, Diff = $\text{Ln}(m_T) - \text{Ln}(m_R)$; s^2 is the intrasubject variance [estimated by the mean square error (MSE) of ANOVA], and N is the number of subjects.

As shown earlier (7), in the case where the upper limit of the 90%CI falls exactly on the upper preset BE limit, Diff becomes equal to Diff_{max} which is the maximum acceptable difference between means:

$$\begin{aligned}\text{Upper limit of the 90\% CI} &= \exp(\text{Diff}_{\text{max}} \pm t_{0.05,N-2}\sqrt{s^2 2/N})\\&= \text{Upper BE limit}\end{aligned} \tag{2}$$

As can be seen from Eq. 2, the maximum acceptable difference, and therefore the maximum acceptable geometric mean ratio GMR_{max} (GMR_{max} + $\exp[\text{Diff}_{\text{max}}]$), for a given number of subjects, is related not only to the estimated intrasubject variance, but also to the value of the preset upper BE limit. Therefore, the maximum difference for the classic preset Upper BE limit, 1.25, is calculated by

$$\begin{aligned}\text{Diff}_{\text{max}} &= \text{Ln(Upper BElimit)} - (t_{0.05,N-2}\sqrt{2/N})s\\&= \text{Ln}(1.25) - (t_{0.05,N-2}\sqrt{2/N})s\end{aligned} \tag{3}$$

Equation 3 clearly shows that the value of Diff_{max} diminishes as the variability increases and consequently BE of HV drugs becomes more difficult to be proven.

On the other hand, if the Upper BE limit is defined as a fixed multiple of intrasubject variability, that is, Upper BE limit = exp(ks), according to the method proposed by Boddy *et al.* (10), Diff_{max} is given by Eq. 4:

$$\text{Diff}_{\text{max}} = ks - (t_{0.05,N-2}\sqrt{2/N})s \tag{4}$$

where k is a proportionality constant. Several values were assigned to k, that is, 1.116 (11), 1 (10), and 0.75 (7). For N = 12 to 36, the quantity $t_{0.05,N-2}\sqrt{2/N}$ varies from 0.74 to 0.40. Therefore, the right hand side of Eq. 4 is positive and shows that the value of Diff_{max} increases with the variability. At high level of variation, Diff_{max} risks to attain a value exceeding the "goal post" of Ln(1.25).

Rationale for the Development of the New BE Limits

An inverse approach focusing on the control of Diff_{max} was used for the development of the new scaled BE limits. In this context, Diff_{max} was fixed to a specific value and then the appropriate value for the Upper BE limit was calculated:

$$\begin{aligned}\text{Diff}_{\text{max}} &= \text{Ln(Upper BE limit)} - (t_{0.05,N-2}\sqrt{2/N})s\\&= k_2\text{Ln}(1.25)\end{aligned} \tag{5}$$

where k_2 is a "constrain" factor for the Diff_{max} value. Equation 5 yields:

$$\text{Ln(Upper BE limit)} = (t_{0.05,N-2}\sqrt{2/N})s + k_2\text{Ln}(1.25) \tag{6}$$

The scaled BE limit defined by Eq. 6 depends not only on the variability (s) but also on the number of subjects participating in the BE study. However, it would be more convenient to define BE limits not dependent on N. This can be accomplished by re-writing Eq. 6 as:

$$\text{Ln(Upper BE limit)} = k_1 s + k_2\text{Ln}(1.25) \tag{7}$$

where k_1 is a proportionality factor. The development of Eq. 7 satisfies the need of i) "constraining" (9) the GMR_{max} accepted values and ii) scaling the BE limits according to a multiple of residual variability (10). Equation 7 can be also viewed as a general form of BE limits, scaled or unscaled. Indeed, if $k_1 = 0$ and $k_2 = 1$, Eq. 7 reduces to the classic definition of unscaled BE limits. On the other hand, if k_1 equals to 1.116 or 1 or 0.75, and $k_2 = 0$, then the resulting Upper BE limits correspond to previously reported scaled BE limits (7,10,11).

The value of k_1 in Eq. 7 can be chosen to control the steepness of the ascending trend of the upper BE limit as variability increases. For a "typical" number of subjects, for example, N = 24, the quantity $t_{0.05,22}\sqrt{2/24}$ equals 0.496. Therefore, a possible choice for k_1 is 0.496. Accordingly, the minimum value of the Upper BE limit is equal to $\exp(k_2\text{Ln}(1.25))$ for the theoretical case s = 0. However, Eq. 5 indicates that $\text{Diff}_{\text{max},n=24}$ is constant and equal to $k_2\text{Ln}(1.25)$. If N < 24, then Diff_{max} shows a descending trend as variability increases. The inverse is observed for N > 24.

A possible choice for the "constrain" factor, which is the simplest one, is $k_2 = 1$. In this case, the resulting scaled BE limits termed BELscN1, Eq. 8, are always wider than the classic unscaled BE limits:

$$\mathrm{Ln(Upper\ BELscN1)} = 0.496s + \mathrm{Ln}(1.25) \tag{8}$$

An inhered problem of the scaled methods is that large differences between the means can be accepted with substantial probabilities. Obviously, the BE limits constructed on the basis of Eq. 8 will have the same drawback. This inhered problem prompted us to re-examine the choice of k_1 in order to design scaled BE limits showing high statistical power under the condition of true BE, and incorporating an effective constraint for GMR.

The simplest approach is to design scaled BE limits by combining only the best performance of the "permissive" BELscN1 under the condition of true BE with GMR = 1 and the performance of classic unscaled BEL when GMR = 1.25. This situation corresponds to $k_1 = 0.496$ when GMR = 1 and $k_1 = 0$ when GMR = 1.25, respectively. The combination of these properties into a single criterion requires a GMR dependent factor k_1, that is, $k_1 = f(\mathrm{GMR})$. This function describes a more general aspect of a constrained criterion. Assuming that k_1 is linearly related to GMR, $k_1 = a + b\mathrm{GMR}$, and the pairs (GMR, k_1) mentioned above with (1, 0.496) and (1.25, 0) satisfy the linear relationship, the following expression for k_1 was derived: $k_1 = (5 - 4\mathrm{GMR})0.496$. These calculations lead to new BE limits (termed hereafter BELscG1) incorporating, both, the intrasubject variability, s, and the GMR, of the specific BE study. Accordingly, the upper limit for BELscG1 is:

$$\mathrm{Ln(Upper\ BELscG1)} = (5 - 4\mathrm{GMR})0.496s + \mathrm{Ln}(1.25) \tag{9}$$

Unscaled BE limits, as mentioned in the Introduction section, allow large differences between the means for drug products with low variability. Therefore, another approach was also undertaken in order to design scaled BE limits with a less permissive behavior at GMR = 1.25. A rather conservative choice for k_1 and k_2 values in Eq. 7 can be derived from the consideration of two multisource drug formulations (T1 and T2) each declared bioequivalent with the same reference product (R) in separate BE studies. Assuming equal number of subjects and equal residual variabilities, as mentioned earlier (7) for an extreme theoretical case, where the upper limit of the 90%CI for T1/R is 1.25 and the lower limit of the 90%CI for T2/R is 0.80, the difference between the two test means equals twice the $\mathrm{Diff}_{\max}$. In order to reduce the maximum difference of means accepted, the values of $k_1 = 0.496 \cdot 0.5$ and $k_2 = 0.5$ were considered and the so derived scaled BE limits are termed hereafter BELscN2; the upper limit for BELscN2 is:

$$\mathrm{Ln(Upper\ BELscN2)} = 0.5[0.496s + \mathrm{Ln}(1.25)] \tag{10}$$

The BELscN2 limits are narrower than the classic unscaled BE limits at low level of variability but they become wider as variability increases. In order to reduce the large differences between the means allowed at a low variability level, the best performance of the "permissive" BELscN1 when GMR = 1 was combined to the more "strict" performance of BELscN2 when GMR = 1.25 The incorporation of these properties into a single criterion was obtained using the (GMR, k_1) pairs with (1, 0.496) and (1.25, 0.5·0.496) and the (GMR, k_2) pairs with (1,1) and (1.25, 0.5), in a similar way to that used for the abovementioned method BELscG1. In this case, the so derived scaled BE limits, termed BELscG2, incorporate GMR-dependent k_1 and k_2 factors. The resulting upper limit for BELscG2 is

$$\mathrm{Ln(Upper\ BELscG2)} = (3 - 2\mathrm{GMR})[0.496s + \mathrm{Ln}(1.25)] \tag{11}$$

It should be noted that Eqs 9 and 11 apply for GMR ≥ 1; when GMR < 1 the reciprocal of GMR is used to calculate the upper BE limit. Starting from Eq. 1, the Lower BE limit for BELscN1, BELscN2, BELscG1 and BELscG2 can be calculated in a similar way.

Scaled BE Limits Considered in the Current Study

Several methods for scaling BE limits reported in the literature (7,9–11), the clasic unscaled BE limits, and the new approaches were evaluated for comparative purposes. Various alternative possibilities for scaled average BE (ABEsc) have been already proposed (11). As mentioned in the Introduction section, a model for scaled average BE (ABEsc) can be readily converted to the corresponding model of scaled BE limits (BELsc). In addition, when investigated, the two approaches yielded very similar results (18). Therefore, only BELsc methods were used in the present investigation.

Based on Eq. 7, the general form of Upper and Lower BE limit can be written as

$$\mathrm{Ln(Upper,\ Lower\ BE\ limit)} = \pm[k_1 s + k_2 \mathrm{Ln}(1.25)] \tag{12}$$

Using the notation of Eq. 12, the BE limits considered in the present study are listed in Table I.

Extreme GMR vs. CV Plots

Extreme values of GMR accepted on the basis of the various scaled BE limits considered were calculated as a function of intrasubject variability (expressed as ANOVA-CV). Assuming two-period crossover BE studies, $\mathrm{GMR}_{\max}$ values

Table I. Methods Based on Scaled BE Limits; k_1 and k_2 are the Factors of Eq. 12

Method	Description	k_1	k_2	Reference
BELscG1	Scaled BE limits incorporating a GMR-dependent constraint criterion	(5–4GMR)0.496	1	This study
BELscG2	Scaled BE limits incorporating a GMR-dependent constraint criterion	(3–2GMR)0.496	3-2GMR	This study
BELscN1	Scaled BE limits incorporating a constant constraint criterion for GMR	0.496	1	This study
BELscN2	Scaled BE limits incorporating a constant constraint criterion for GMR	0.5 · 0.496	0.5	This study
BEL	Unscaled BE limits	0	1	(2)
BELsc1	Scaled BE limits	1.116	0	(11)[a]
BELsc2	Scaled Be limits	1.000	0	(10)
BELsc3	Scaled BE limits	0.759	0	(7)
BELsc1M	Mixed model: Unscaled BE limits up to CV 20% and BELsc1 for CVs >20%	0 or 1.116	1 or 0	(11)[a]
BELsc2C	BELsc2 with the additional criterion: 0.80 ≤ GMR ≤ 1.25	1.000	0	(9)[a]

[a] The corresponding ABEsc procedure is reported.

can be computed by substituting in Eq. 13 the appropriate Upper BE limit for each procedure considered

$$\begin{aligned} GMR_{max} &= \exp(Diff_{max}) \\ &= \exp[Ln(\text{Upper BE limit}) - (t_{0.05,N-2}\sqrt{2/N})s] \end{aligned} \quad (13)$$

The corresponding minimum values of GMR (GMR_{min}) can be computed in a similar way. Extreme values of GMR were calculated for several levels of sample size, that is, for N = 12, 16, 18, 24, 32, and 36. Extreme values of GMR accepted by BELscG1 and BELscG2 were computed numerically with an iterative method. GMR vs. CV plots were constructed and used as a tool for the assessment of the different scaled approaches in bioequivalence studies.

Simulations

Two-treatment, two-period, crossover bioequivalence studies, with equal number of subjects in each sequence, were simulated and evaluated using the BE limits listed in Table I. BE was declared in each simulated crossover study if the 90%CI around the ratio of the estimated geometric means (GMR) for the 2 drug products was between preset BE limits; 12, 24, or 36 subjects were assumed to participate in the simulated trials. Log-normally distributed parameters were assumed. The true CV values considered for the simulations, ranged from 10% to 40%. The standard deviations (σ) of the logarithmically transformed parameters were calculated from the preset CV by $\sigma = \sqrt{Ln(1 + CV^2)}$. The average parameter value for the reference formulation was set to 100 arbitrary units. The true ratio of geometric means was gradually changed, from the condition of true BE to increasing deviations from BE. Therefore, simulated GMR values ranged from 1.00 to 1.50.

Twenty thousand simulated BE trials were performed under each condition. The percentage of simulated studies in which BE was accepted was then recorded. Power curves were constructed by plotting the percentage of acceptance *vs.* the true value of the GMR. The conditions of clinical BE trials were simulated by developing a computer program in Fortran. The program was validated by comparing some of the simulated acceptances of BE studies using BEL with previously published power curves (11,18,19). In addition, the overall accuracy of the simulation method was assessed by recording the number of times the true GMR value is within the 90% confidence interval. In all cases, the percent of true GMR value within the confidence interval was found to be 90%.

RESULTS AND DISCUSSION

BE Limits-CV and GMR-CV Relationships for the Methods Considered

Figure 1 presents a graphical illustration of BE limits as a function of residual (intrasubject) variability, expressed as ANOVA-CV %. The two new approaches BELscG1 and BELscG2 are shown along with the classic unscaled BEL of 0.80–1.25 and two scaled methods BELsc1M [the interesting mixed model, (11)] and BELsc2 [a typical scaled procedure, (10)]. Because both BELscG1 and BELscG2 also vary with the magnitude of GMR, the shaded areas of Fig. 1 indicate the range of GMR values used, namely 0.80 ≤ GMR ≤ 1.25. As can be seen from Fig. 1, all scaled BE limits, in contrast to the classic invariant BEL, become wider as variability increases. However, BE limits widening for BELscG1 and BELscG2 has less steep slopes than the other scaled methods.

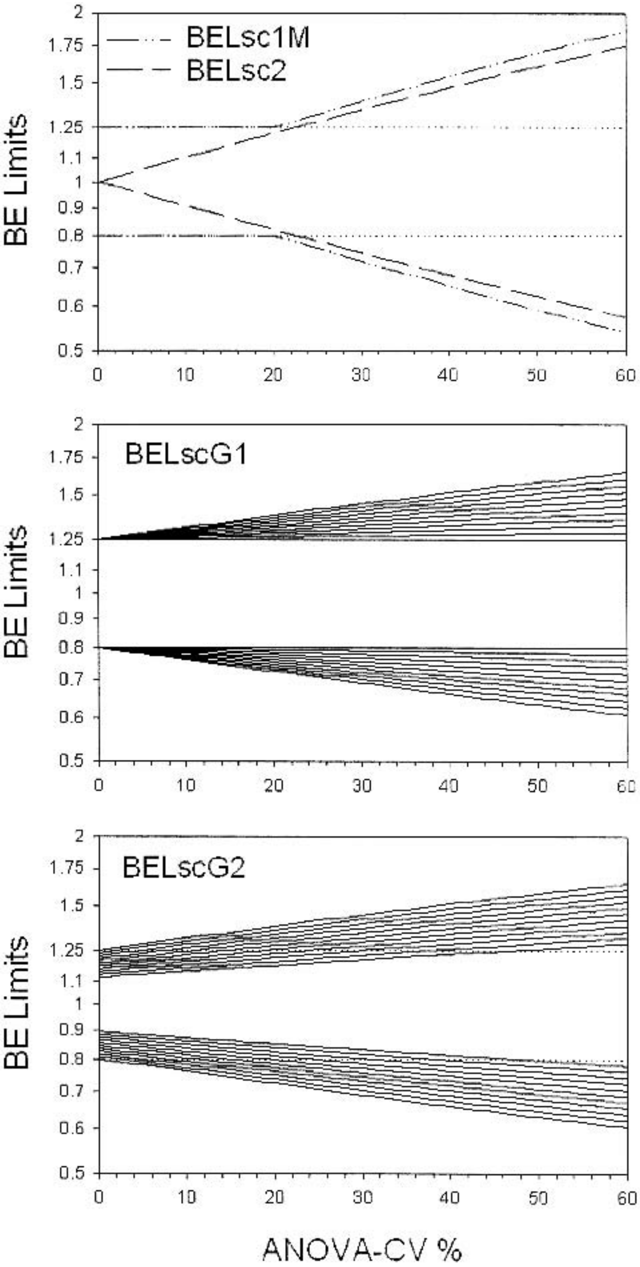

Fig. 1. BE limits as a function of intrasubject variability (ANOVA-CV %) for five methods are shown. The BE limits for the methods BELscG1 and BELscG2 vary with the magnitude of GMR; the two shaded areas correspond to the GMR range 0.80–1.25 studied. The borderlines of these areas correspond to GMR values (from top to bottom): 1, 1.25, 0.80, 1. The dotted lines correspond to the classic unscaled BEL.

Figure 2 shows the maximum and minimum GMR accepted values for two-period crossover bioequivalence studies, by various scaled BE approaches, as a function of the residual variability (intrasubject ANOVA-CV). Several levels of sample size were considered, namely, N = 12, 16, 18, 24, 32,

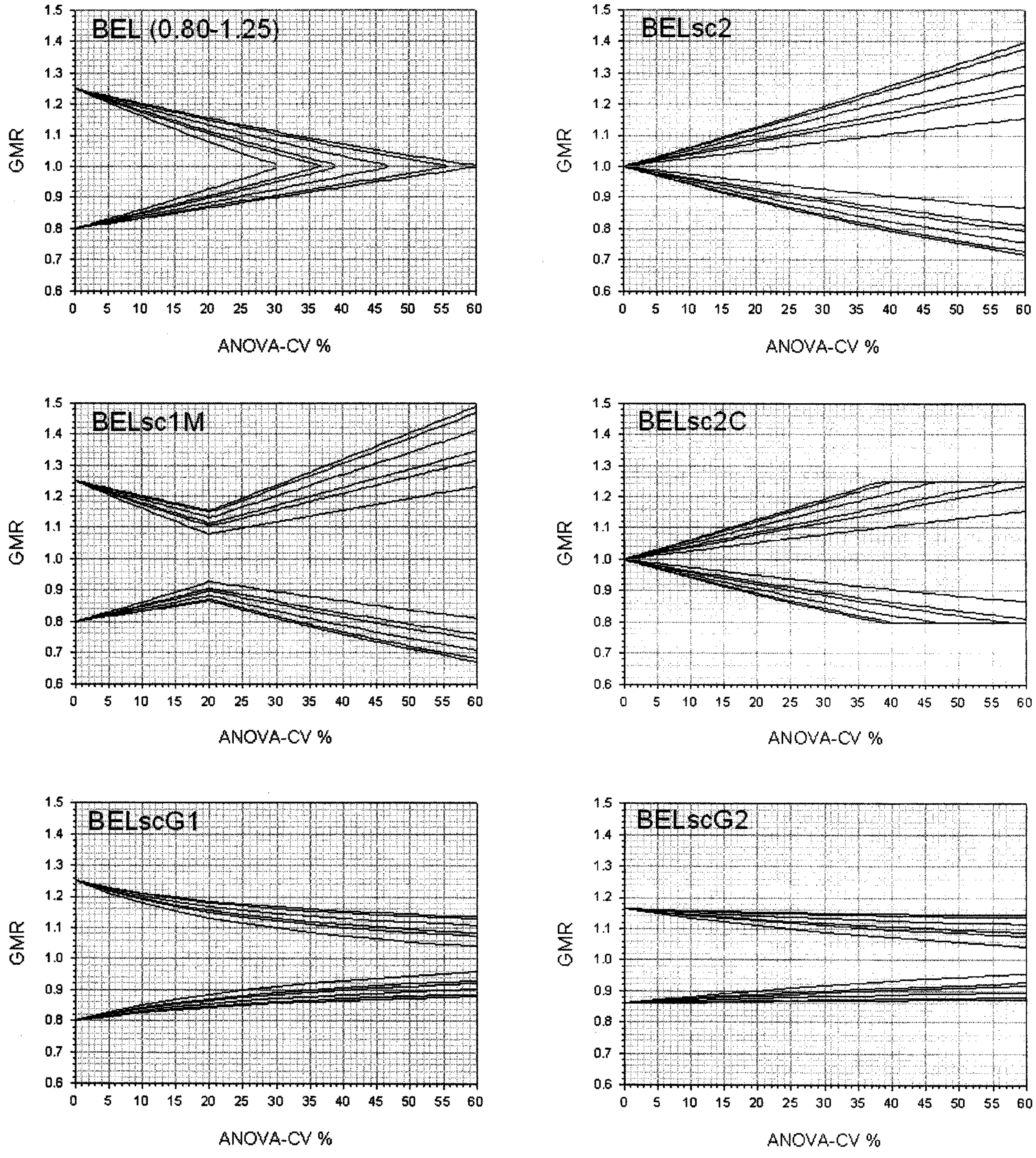

Fig. 2. Extreme GMR (values) accepted by six methods as a function of intrasubject variability (ANOVA-CV). Each line corresponds to a different level of sample size, N, considered. Key (from top to bottom): N = 36, 32, 24, 18, 16, and 12, for the upper family of curves (GMR_{max} lines) and N = 12, 16, 18, 24, 32, and 36 for the lower family of curves (GMR_{min} lines).

and 36. Extreme values of GMR for the classic unscaled BEL 0.80–1.25 were also calculated for comparison. For a given sample size, upper and lower curves, in all graphs of Fig. 2, correspond to the maximum acceptable deviation of GMR from unity, when test and reference formulations are declared bioequivalent. Therefore, all values of GMR lying between the minimum (GMR_{min}) and the maximum (GMR_{max}) line, for the given sample size, indicate bioequivalence between the two drug products.

As expected, unscaled BEL allows large deviations between the means, that is, allows great deviations from unity of the GMR values, for drug products with low residual variability. On the other hand, unscaled BEL, appears to be very strict for HV drugs since the range of the GMR accepted values decreases dramatically for CV values higher than 20%. This observation agrees with the very well known fact that it is very difficult to prove bioequivalence when the intrasubject variability is high. As can be seen in Fig. 2, when CV >30%, the maximum difference allowed for GMR, using the unscaled BEL, is less than 10%, even when a large number of subjects is used. For example, at CV = 40% the maximum GMR accepted for the two products is only 1.03 and 1.07 assuming 24 and 36 subjects participating in the BE study, respectively.

To overcome these difficulties several methods for expanding BE limits have been proposed. A typical example of the most widely used scaled criterion (10), (BELsc2, Table I) is presented in Fig. 2. As can be seen, when variability is low, very small deviations of GMR from unity are permitted. For example, at CV = 10% the allowed GMR_{max} value is only 1.03 when 12 subjects are assumed to participate in the BE study. Moreover, even with a large number of subjects, N = 36, the GMR_{max} allowed is only 1.06. Consequently, BELsc2 appears to be very strict for drugs with low variability and probably inappropriate even for the evaluation of drugs with narrow therapeutic range. This also applies for other scaling methods, for example, BELsc3 (20). As variability increases, BELsc2 becomes very liberal, allowing GMR_{max} greater than 1.25. For example, at high level of variation, that is, CV = 40%, the GMR_{max} value allowed is 1.26 when 36 subjects are assumed to participate in the BE study. In this case, the corresponding upper limit of the 90% confidence interval is 1.46. At this high level of variability, the scaling method BELsc1, Table I, is even more liberal than BELsc2 (data not shown). The joint application of BELsc2 with a constraint on GMR between 0.80 and 1.25 (9) (method BELsc2C, Table I) is not too liberal in terms of GMR for HV drugs, but it is too strict for low variability drugs as mentioned above for all scaled methods. According to Fig. 2, when the mixed model BELsc1M is used, the range GMR_{max}–GMR_{min} decreases as CV values increase up to 20%. For CV = 20% the range GMR_{max}–GMR_{min} reaches a minimum value and then increases again for values of CV higher than 20%. Consequently, this approach is less permissive for drugs with moderate variability than for drugs with low or high variability. This discontinuity in monotony of the GMR vs. CV plots might be an unfavorable property of the method, because it appears to "punish" drugs with moderate variability.

The two new proposed scaling methods are shown on the bottom graphs of Fig. 2. The general aspect of the GMR vs. CV plots is similar for the two approaches. Both new scaled methods become more strict as variability increases, in the same sense as unscaled BEL, but with less steep curves. As expected, based on the design of these scaled methods, BELscG1 appears to be more liberal than BELscG2 at low variability level. For the theoretical case of no variability, BELscG1 allows, as the classic unscaled BEL, GMR_{min} and GMR_{max} values equal to 0.80 and 1.25, respectively. In contrast, at CV = 0%, BELscG2 allows GMR_{min} and GMR_{max} values of 0.86 and 1.16, respectively. For both approaches there is a gradual decrease of the range GMR_{max}–GMR_{min} as variability increases and the two methods become practically identical at high variability level, that is, at CVs roughly over 35%.

Similar plots were constructed (data not shown), relating extreme GMR values allowed with Mean Square Error (MSE) from ANOVA. The GMR vs. CV plots of Fig. 2 or alternatively GMR vs. MSE plots can be used directly for the assessment of bioequivalence using the different approaches examined. In this context, two drug products evaluated in a two-period crossover study with a given number of subjects N, can be declared bioequivalent when the (CV, GMR) datum point based on the estimates of the study lies between the minimum (GMR_{min}) and the maximum (GMR_{max}) values.

We also examined the methods BELscG1 and BELscG2, in comparison to the classic BEL, from the point of view of the "sensitivity" (∇) proposed by Shuirmann (21). The results of the analysis revealed that at high ANOVA-CV % the two new methods, present higher ∇ values (data not shown) than the unscaled BEL.

Assessment of BE Using the Methods Considered

Table II presents the percentages of studies in which BE was accepted by applying the classic unscaled BEL and the various scaled methods listed in Table I. Two-period crossover simulated studies performed with 24 subjects and intrasubject coefficient of variation of 30% were assumed. The simulated ratio of the geometric means (GMR) varied from 1.00 to 1.45. As expected, the unscaled BEL procedure yields quite low acceptances, and therefore low statistical power

Table II. Acceptance (%) of Bioequivalence of Two Drug Products in Two-Period Crossover Simulated Studies with 24 Subjects (CV = 30%) Using the Methods Listed in Table I

	GMR[a]									
Method	1.00	1.05	1.10	1.15	1.20	1.25	1.30	1.35	1.40	1.45
BELscG1	89.8	85.3	69.1	49.2	30.1	16.1	7.1	3.0	1.1	0.3
BELscG2	86.3	81.8	63.6	43.2	24.9	12.6	5.2	2.0	0.7	0.2
BELscN1	99.2	98.0	93.5	83.7	68.9	49.9	32.2	18.1	9.0	3.9
BELscN2	35.9	30.7	20.3	10.1	4.3	1.5	0.5	0.2	0.0	0.0
BEL	63.7	56.5	40.2	23.8	11.6	5.0	1.8	0.6	0.2	0.0
BELsc1	95.7	92.6	82.9	67.6	48.9	31.5	17.9	8.9	3.9	1.5
BELsc2	90.6	85.5	71.5	52.9	33.8	19.1	9.3	4.0	1.4	0.5
BELsc3	63.1	55.4	39.2	22.3	10.6	4.3	1.4	0.5	0.1	0.0
BELsc1M	95.8	92.6	83.0	67.7	49.0	31.6	18.0	8.9	3.9	1.5
BELsc2C	90.6	85.5	71.5	52.9	33.8	19.1	9.3	4.0	1.4	0.5

[a] The true ratio of geometric means.

(e.g., 63.7% when true GMR = 1). The scaled procedure BELsc3 also yields similar low acceptances, and appears to be the most "strict" among the proposed scaled methods in the literature. The proportion of accepted studies is substantially higher when BELsc2 and BELsc2C are considered. For GMR = 1, the statistical power reaches the fairly acceptable level of 0.906. At this level of variation (CV = 30%), the performance of BELsc2C is practically identical to that of BELsc2. The constraint on GMR has no effect on the acceptance of bioequivalence, even at high values of the true ratio of geometric means. Both BELsc2 and BELsc2C, appear to be somewhat permissive when the two drug products differ more than 25%. The two methods show an acceptance percentage of 19.1%, when true GMR = 1.25. The performance of BELsc1M is very similar to that of BELsc1. When there is no difference between the two products (GMR = 1), the percentages of accepted BE studies are 95.8% and 95.7%, respectively. However, the two methods become very permissive, even when GMR = 1.25, with the percentage of accepted studies reaching the rather high level of 31.6%. As expected from their design, BELscN1 appears to be too "liberal" while BELscN2 appears to be very "strict", at all levels of the true GMR assumed. It is interesting to mention that the highest statistical power is observed for BELscN1 when GMR = 1, and the lowest percentage of acceptance is observed for BELscN2 when GMR = 1.25. The new methods BELscG1 and BELscG2, show almost similar proportion of acceptances, that is, 89.8% and 86.3% when GMR = 1, but they appear to be less permissive (16.1% and 12.6% when GMR = 1.25) than BELsc1, BELsc2, BELsc2C, and BELsc1M.

We further evaluated all procedures using simulated data assuming N = 12, 24, and 36 and CV = 10%, 20%, 30%, and 40%. The results presented in Fig. 3 refer only to six of the ten methods listed in Table I, namely, the classic unscaled BEL, the most typical scaled method BELsc2 (10), the two recently proposed scaled procedures BELsc1M (11) and BELsc2C (9), and the two new methods BELscG1 and BELscG2 developed in this study. Figure 3 shows the percentage of BE studies accepted at increasing true ratios of the geometric means (GMR), assuming two-period simulated crossover studies with 24 subjects.

As can be seen in Fig. 3, at low variation (CV = 10%) the unscaled BEL, and the methods BELsc1M, BELscG1 and BELscG2 show similar performance, with 100% acceptance when GMR = 1. Nevertheless, the new approach BELscG2 exhibits a slightly steeper power curve. The scaled procedures BELsc2 and BELsc2C yield substantially lower acceptance and appear very strict, since they show 0.1% acceptance at GMR = 1.15. These results are in full agreement with the extreme GMR *vs.* CV plots of Fig. 2. The GMR_{max} values at CV = 10% for BEL, BELsc1M, BELscG1, and BELscG2 methods vary from 1.15 to 1.20, while the GMR_{max} allowed for BELsc2 and BELsc2C is only 1.05. When an intermediate level of variation is assumed (CV = 20% and GMR = 1), BELscG1 and BELscG2 exhibit 99.7% and 99.0% of acceptance, respectively, while BELsc1M and unscaled BEL show slightly lower percentages, that is, 98.2% and 96.8%, respectively. Lower statistical power (0.906) is observed for BELsc2 and BELsc2C. When GMR = 1.25, BELscG1 is found to be the most permissive among the methods, showing 12.9% of acceptance. This can be also explained by the GMR *vs.* CV plots of Fig. 2 because at CV = 20% the highest GMR_{max} value (1.17) is observed for the method BELscG1. BELscG2 exhibits a more strict behavior at GMR = 1.25 compared to

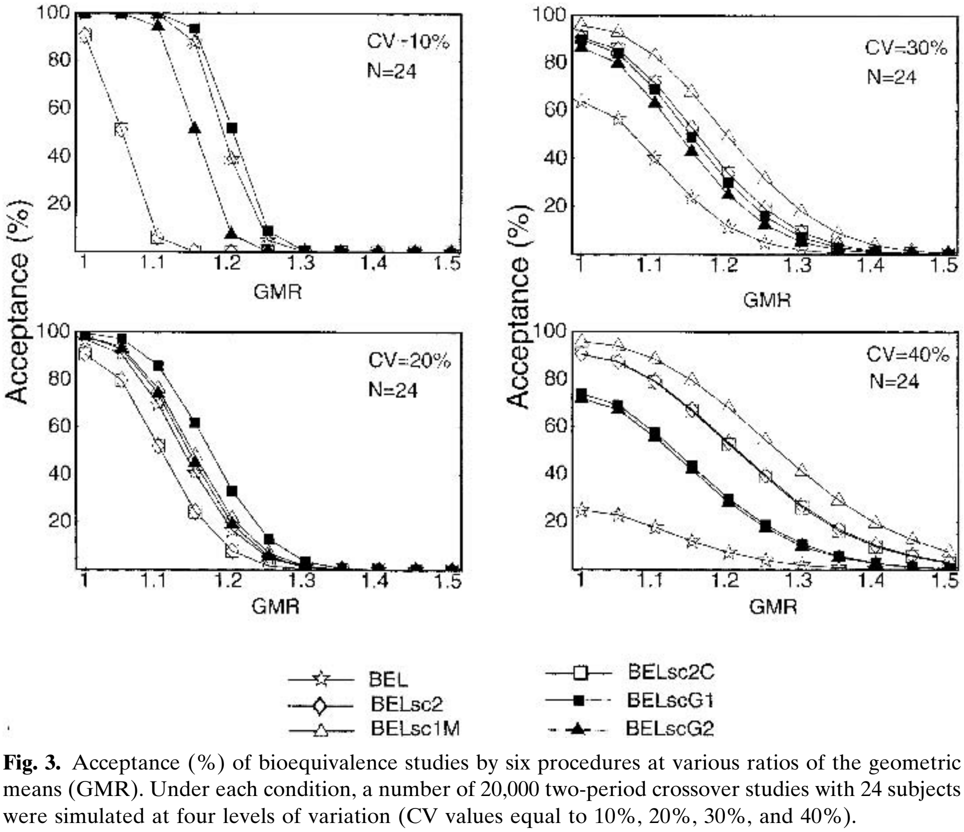

Fig. 3. Acceptance (%) of bioequivalence studies by six procedures at various ratios of the geometric means (GMR). Under each condition, a number of 20,000 two-period crossover studies with 24 subjects were simulated at four levels of variation (CV values equal to 10%, 20%, 30%, and 40%).

BELscG1 and BELsc1M, since it shows a fairly low acceptance of 5.6%.

As variation increases (e.g., CV = 30% and 40%) all scaled procedures yield much higher acceptance than the classic unscaled BEL, Fig. 3. At CV = 30% and GMR = 1, the proportion of accepted BE studies ranges from 86.3% to 95.8% for the scaled procedures considered in this study, while only 63.7% of the studies are accepted by the unscaled BEL. At CV = 40% and GMR = 1, the statistical power of BELsc2C, BELsc2 and BELsc1M continues to be high (ranging from 0.904 to 0.957) and fairly high for the new approaches BELscG1 and BELscG2 (0.737 and 0.719, respectively). The well-known inability of the classic unscaled BEL to demonstrate bioequivalence when really exists, is reflected by the very low proportion of BE studies accepted (24.7%). Although scaled procedures at high variations show higher statistical power, compared to the unscaled BEL, unfortunately, they also appear to be more permissive as the ratio of geometric means increases. This is an unfavorable performance of the previously published scaled methods. When GMR = 1.25, at CV = 30%, BELsc2 and BELsc2C show 19.1% of acceptance, while BELsc1M shows 31.6% of acceptance. The new methods BELscG1 and BELscG2 show a relatively low percentage of acceptance (16.1 and 12.6%, respectively). At CV = 40%, the scaled methods become extremely permissive, with 38.9%, 39.6%, and 54.5% of acceptance for BELsc2C, BELsc2, and BELsc1M, respectively, Fig. 3. The additional constraint on GMR used for BELsc2C appears to have only a minor effect on the acceptance of bioequivalence, under these simulated conditions. At this high level of variation, the methods BELscG1 and BELscG2 show rather "reasonable" acceptances of 18.8% and 17.6%, respectively, when the true GMR = 1.25. The aforementioned results of the power curves are also supported by the GMR vs. CV plots of Fig. 2. When 24 subjects are assumed to participate in the BE study, GMR_{max} values at CV = 30% for BELscG1 and BELscG2 procedures are 1.15 and 1.13, respectively. The GMR_{max} values for BELsc2 and BELsc1M are higher, that is, 1.16 and 1.20, respectively. At higher level of variation, CV = 40%, the GMR_{max} value for both BELscG1 and BELscG2 is only 1.13 while for BELsc2 and BELsc1M the GMR_{max} values are 1.21 and 1.27 respectively, explaining the very permissive performance of these methods.

Overall, the two new approaches BELscG1 and BELscG2 exhibit a good performance when GMR = 1, showing a reasonably high statistical power for CV values 30% and 40%, Fig. 3. On the other hand, both BELscG1 and BELscG2 show the lowest percentages of acceptance among the scaled methods when GMR = 1.25, that is, 16.1%, 12.7% for CV = 30% and 18.8%, 17.6% for CV = 40%, respectively. Undoubtedly, the performance of the new methods BELscG1 and BELscG2 is by far better than the previously published scaled methods at high GMR values.

The method of scaling BE limits was mainly proposed for the assessment of bioequivalence of HV drugs, usually evaluated with a large number of subjects. Nevertheless, a very interesting approach has been mentioned earlier about the potential utility of scaling for the evaluation of toxic drugs (7). Notably, it has been mentioned that scaling could be advantageous in the case of toxic drugs with low variability and narrow therapeutic ranges. Therefore, the results obtained, considering 2-period studies with a small number of subjects, for drugs with low variability, are of relevant importance. At low level of variation (CV = 10%, data not shown), when 12 subjects are assumed to participate in the simulated crossover BE studies, a very high producer risk is observed for the scaled methods BELsc1, BELsc2, BELsc2C, and BELsc3. This constitutes an unfavorable performance of the aforementioned scaled methods, and consequently, they are rather inappropriate for the evaluation of toxic drugs with low variability and narrow therapeutic range. On the other hand, the new approach BELscG2 exhibits similar performance, with unscaled BEL and BELscG1 (with 100% of acceptance at GMR = 1), but BELscG2 also shows slightly steeper power curves, resulting in a more "strict" criterion as the difference between the means becomes higher. Therefore, BELscG2 could be somewhat preferable for the evaluation of toxic drugs.

Figure 4 shows the percentage of BE studies accepted by the same methods used for the simulations of Fig. 3 assuming a small (N = 12) and a large (N = 36) number of subjects. Thus, the results reported in Fig. 4 refer to two-period simulated crossover trials with 12 and 36 subjects at two levels of CV (20%, 30% for N = 12 and 30%, 40% for N = 36).

The top left graph of Fig. 4, illustrates the results obtained from the simulation of a rather rare case, where only 12 subjects participate in BE studies of a drug presenting moderate level of variation (CV = 20%). As can be seen from Fig. 4, when GMR = 1 there are clear differences of the percentage of acceptance among the various investigated methods. The methods BELsc2 and BELsc2C show similar results with those obtained at low level of variability, CV = 10%, presenting the lower proportion of BE studies accepted (46.4%). The classic unscaled BEL shows much higher proportion of accepted BE studies (64.7%) and the mixed model BELsc1M, exhibits a better performance with 75.8% of acceptance. The new method BELscG2 shows even higher percentage of acceptance (80.4%), reaching an adequate level of statistical power, while the new approach BELscG1 presents the best performance with 87.2% of acceptance. The results obtained from an extreme hypothetical case of a HV drug evaluated in BE studies with 12 subjects are presented at the bottom left graph of Fig. 4. At this high level of variation (CV = 30%), lower proportions of accepted studies are generally observed. Under the condition of true bioequivalence, as expected, the classic unscaled BEL, has very low (16.2%) percentage of acceptance, while the scaled methods, designed to be more liberal as variability increases, show higher proportions of accepted studies, namely, 46.4, and 63.7% for BELsc2, and BELsc1M, respectively. The two new approaches BELscG1 and BELscG2 exhibit a good performance. Even under this extreme scenario, the statistical power when GMR = 1, reaches a fairly high level of 57.9 and 53.0% of accepted BE studies, respectively. When GMR = 1.25, the proportions of accepted BE studies by the scaled methods range from 10.2% to 17.0%. The percentages of acceptance for the new methods BELscG1 and BELscG2 are 14.8% and 12.7%, respectively.

The two graphs on the right-hand side of Fig. 4 illustrate the results obtained for HV drugs in crossover studies with 36 subjects. Under these conditions, all scaled methods show very high proportions of accepted studies when there is no true difference between the two means. At CV = 30% the percentages of acceptance are 98.1% to 99.6%, while they are

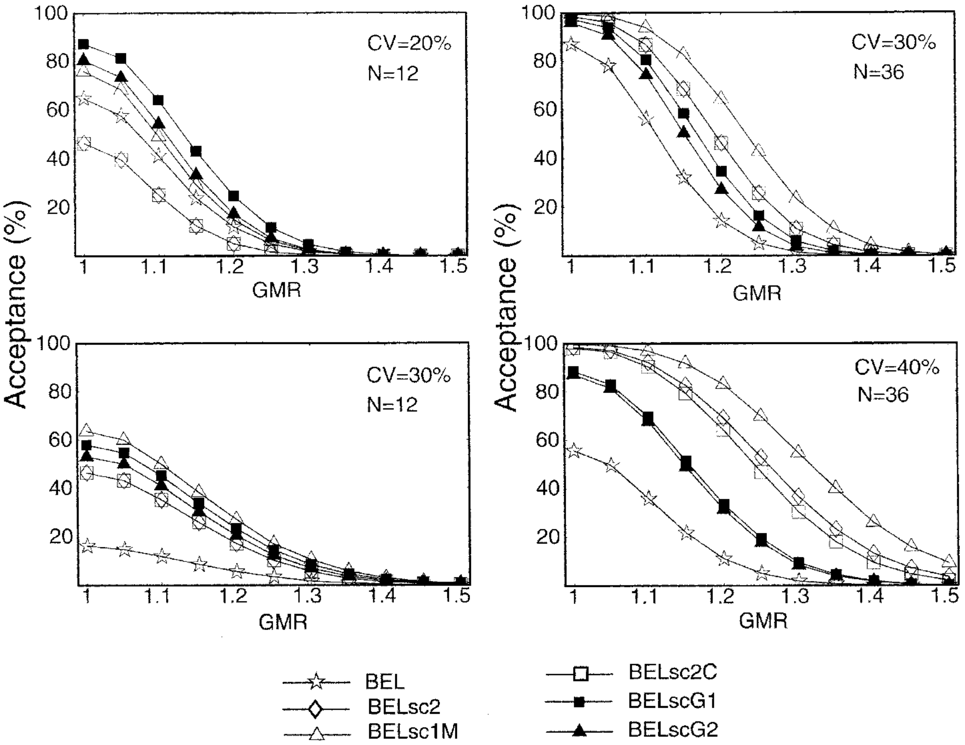

Fig. 4. Acceptance (%) of bioequivalence studies by six procedures at various ratios of the geometric means (GMR). Under each condition, a number of 20,000 two-period crossover studies with 12 or 36 subjects were simulated at two levels of CV (20%, 30% when N = 12, and 30%, 40% when N = 36).

slightly lower as variability rises, that is, 93.7% to 99.6% at CV = 40%. In contrast, the unscaled BEL exhibits a poor performance, even with this large number of subjects, presenting a statistical power of 0.870 and 0.557 for CV 30% and 40%, respectively. It is also worthy to mention that the analysis performed by calculating the "sensitivity," ∇ (21), of the methods BELscG1, BELscG2, and the classic BEL for N = 12, 24, 36, and CV = 20%, 30%, 40% revealed (data not shown) that the two new methods require a smaller N to prove BE at high ANOVA-CV % and exhibit comparable or even higher "sensitivity" than the unscaled BEL. These results are in full agreement with the power curves presented in Figs. 3 and 4. Moreover, when evaluating scaled methods, especially at high variability level, it is also very important to examine the proportion of accepted BE studies when the true GMR is high. As can be seen from Fig. 4, the methods BELsc2C, BELsc2, and BELsc1M show rather gently declining power curves, resulting in substantial proportions of accepted BE studies, even when true difference between the means exceeds 25%. For example, when GMR = 1.25 and CV = 40%, the percentage of acceptance is 46.7% and 53.0% for BELsc2C, and BELsc2, respectively. The percentage of acceptance for BELsc1M is even larger, and reaches the extremely high proportion of 69.9%. In contrast, the new approaches BELscG1 and BELscG2 present steeper power curves and consequently result in less permissive conclusions. When the true difference between the means is 25%, the lowest acceptances are observed for BELscG1 and BELscG2: 16.5% and 11.6% at CV 30%, as well as 19.4% and 17.6% at CV 40%, respectively.

The need for a large number of subjects is a common problem in BE studies of HV drugs. Scaled methods, overcome this difficulty in part by reducing the total number of subjects required while reaching a high statistical power under the condition of true bioequivalence. On the other hand, the increase in the number of subjects results in a substantial increase of the statistical power of the classic unscaled BEL, while the consumer risk is practically invariant (i.e. 5%). However, for the scaled methods proposed in the literature, the benefit of the lower producer risk is counterbalanced by the substantial increase of the proportions of acceptance, when there is a 25% difference between the means. These observations can be confirmed if one compares the power curves for 24 and 36 subjects in Figs. 3 and 4. For example, under the condition of true BE at CV = 40%, there is indeed a small increase in statistical power (from 95.7% to 99.6%) for the method BELsc1M, when a larger number of subjects is assumed to participate in the study (36 instead of 24). However, BELsc1M results also in more permissive conclusions (from 54.5% to 69.9%) when a 25% difference between the means is assumed. On the contrary, the increase in the number of subjects results in the increment of statistical power of the new scaled methods BELscG1 and BELscG2 when GMR = 1, while only a negligible increase of the permissiveness is observed when there is a 25% difference between the means. For example, when GMR = 1, BELscG2 exhibits an increase in statistical power from 71.9% to 87.2% assuming 24 and 36 subjects, respectively, while the proportion of acceptance remains practically invariant (17.6%) and the lowest observed for the scaled methods considered in this study when GMR = 1.25. Overall, the new proposed scaled methods BELscG1 and BELscG2, offer the possibility of increasing the statistical power, when GMR = 1, by increasing the number of subjects, while keeping the proportions of acceptance, when GMR =

1.25, practically invariant and at the lowest level observed for the scaled methods considered in this study.

CONCLUSIONS

Compared to the two recently proposed scaled procedures, that is, BELsc1M and BELsc2C, the new methods BELscG1 and BELscG2 exhibit better performance. These new approaches appear to be highly effective at all levels of variation investigated. When variability is low, BELscG1 and BELscG2 do not present the drawbacks of the typical scaled methods that tend to be strict, showing increased producer risk, but they exhibit the highest statistical power among the scaled methods investigated. On the other hand, the new approach BELscG2 exhibits similar performance, with unscaled BEL (with 100% of acceptance at GMR = 1), but also shows slightly steeper power curves, resulting in a more "strict" criterion as the difference between the means becomes higher. Therefore, BELscG2 could be somewhat preferable than BEL for the evaluation of toxic drugs. At high levels of variation, BELscG1 and BELscG2 overcome the well-known problems of the unscaled BEL, as they require a smaller number of subjects to prove BE, but also the shortcomings of other scaled methods that tend to be too permissive even when GMR exceeds 1.25. Therefore, especially when variability is high, the new approaches BELscG1 and BELscG2 exhibit a very favorable performance, presenting high statistical power under the condition of true bioequivalence as well as the lowest percentage of acceptance among all the scaled methods examined when the true difference between the means exceeds 25%.

ACKNOWLEDGMENTS

The authors wish to thank the reviewers for their constructive comments.

REFERENCES

1. European Agency for the Evaluation of Medicinal Products. *Note for Guidance on the Investigation of Bioavailability and Bioequivalence.* Committee for Proprietary Medicinal Products (CPMP), London, 2001.
2. Food and Drug Administration. *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Consideration*, Center for Drug Evaluation and Research (CDER), Rockville, MD, 2000.
3. H. Blume and K. Midha. Report of consensus meeting: Bio-international'92, Conference on Bioavailability, Bioequivalence and Pharmacokinetics studies, Bad Homburg, Germany, 20-22 May 1992. *Eur. J. Pharm. Sci.* **1**:165–171 (1993).
4. H. Blume, I. McGilveray, and K. Midha. Report of consensus meeting: Bio-international'94, Conference on Bioavailability, Bioequivalence and Pharmacokinetics studies, Munich, Germany, 14-17 June 1994. *Eur. J. Pharm. Sci.* **3**:113–124 (1995).
5. V. Shah, A. Yacobi, W. Barr, L. Benet, D. Breimer, M. Dobrinska, L. Endrenyi, W. Fairweather, W. Gillespie, M. Gonzales, J. Hooper, A. Jackson, L. Lesko, K. Midha, P. Noonan, R. Patnaik, and R. Williams. Evaluation of orally administered highly variable drugs and drug formulations. *Pharm. Res.* **13**:1590–1594 (1996).
6. S. Anderson and W. Hauck. The transitivity of bioequivalence testing. Potential for drift. *Int. J. Clin. Pharmacol. Ther.* **34**:369–374 (1996).
7. K. Midha, M. Rawson and J. Hubbard. Bioequivalence: switchability and scaling. *Eur. J. Pharm. Sci.* **6**:87–91 (1998).
8. H. Blume, M. Elze, H. Potthast, and B. Schug. Practical strategies and design advantages in highly variable drug studies: multiple dose and replicate administration design. In H.H. Blume and K. Midha (eds.), *Bio-international '92: Bioavailability, Bioequivalence, and Pharmaokinetic Studies.* Medpharm, Stuttgart, 1995, pp. 117–122.
9. L. Tothfalusi, L. Endrenyi, and K. Midha. Scaling or wider bioequivalence limits for highly variable drugs and for the special case of C_{max}. *Int. J. Clin. Pharmacol. Ther.* **41**:217–225 (2003).
10. A. Boddy, F. Snikeris, R. Kringle, G. Wei, J. Oppermann, and K. Midha. An approach for widening the bioequivalence acceptance limits in the case of highly variable drugs. *Pharm. Res.* **12**:1865–1868 (1995).
11. L. Tothfalusi and L. Endrenyi. Limits for the scaled average bioequivalence of highly variable drugs and drug products. *Pharm. Res.* **20**:382–389 (2003).
12. R. Schall and H. Luus. On population and individual bioequivalence. *Stat. Med.* **12**:1109–1124 (1993).
13. R. Patnaik, L. Lesko, M.-L. Chen, and R. Williams. Individual bioequivalence: new concepts in the statistical assessment of bioequivalence metrics. *Clin. Pharmacokin.* **33**:1–6 (1997).
14. K. Midha, M. Rawson, and J. Hubbard. Individual and average bioequivalence of highly variable drugs and drug products. *J. Pharm. Sci.* **86**:1193–1197 (1997).
15. L. Endrenyi, G. Amidon, K. Midha, and P. Skelly. Individual bioequivalence: attractive in principle, difficult in practice. **15**: 1321–1325 (1998).
16. Food and Drug Administration. *Statistical Approaches to Establishing Bioequivalence*. Center for Drug Evaluation and Research (CDER), Rockville, MD (2001).
17. R. Schall. A unified view of individual, population, and average bioequivalence. In. H.H. Blume and K.K. Midha (eds.), *Bio-International '92: Bioavailability, Bioequivalence, and Pharmacokinetic Studies*, Medpharm, Stuttgart, 1995, pp. 91–106.
18. L. Tothfalusi, L. Endrenyi, K. Midha, M. Rawson, and J. Hubbard. Evaluation of the bioequivalence of highly-variable drugs and drug products. *Pharm. Res.* **18**:728–733 (2001).
19. E. Diletti, D. Hauschke, and V. W. Steinijans. Sample size determination: extended tables for the multiplicative model and bioequivalence ranges of 0.9 to 1.11 and 0.7 to 1.43. *Int. J. Clin. Pharmacol. Ther. Toxicol.* **30**:S59–S62 (1992).
20. V. Stakias and M. Symillides. Bioequivalence of generic drug products and modification of bioequivalence limits. *Eur. J. Drug Metab. Pharmacokin* **28**:7–8 (2003).
21. D. J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokin. Biopharm.* **15**:657–680 (1987).