Biometrical Journal **44** (2002) 8, 969–981

Sample Size Requirements for Evaluation of Bridging Evidence

JEN-PEI LIU[1,2*], HUEYMIIN HSUEH[3] and JAMES J. CHEN[4]

[1] Department of Statistics, National Cheng-kung University, Tainan, Taiwan
[2] Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, Taiwan
[3] Department of Statistics, National Cheng-Chi University, Taipei, Taiwan
[4] Division of Biometry and Risk Assessment, National Center for Toxicological Research, Food and Drug Administration, Jefferson, Arkansas 72079, U.S.A.

Summary

This paper addresses issues concerning methodologies on the sample size required for statistical evaluation of bridging evidence for a registration of pharmaceutical products in a new region. The bridging data can be either in the Complete Clinical Data Package (CCDP) generated during clinical drug development for submission to the original region or from a bridging study conducted in the new region after the pharmaceutical product was approved in the original region. When the data are in the CCDP, the randomized parallel dose-response design stratified to the ethnic factors and region will generate internally valid data for evaluating similarity concurrently between the regions for assessment of the ability of extrapolation to the new region. Formula for sample size under this design is derived. The required sample size for evaluation of similarity between the regions can be at least four times as large as that needed for evaluation of treatment effects only. For a bridging study conducted in the new region in which the data of the foreign and new regions are not generated concurrently, a hierarchical model approach to incorporating the foreign bridging information into the data generated by the bridging study is suggested. The sample size required is evaluated. In general, the required sample size for the bridging trials in the new region is inversely proportional to equivalence limits, variability of primary endpoints, and the number of patients of the trials conducted in the original region.

Key words: Extrapolation; Similarity; Equivalence limit; Hierarchical Model.

1. Introduction

For marketing approval of a pharmaceutical product, sponsors are required to provide substantial evidences of the effectiveness and safety from adequate and well-controlled clinical trials. On the other hand, to evaluate the reproducibility of evidences, usually at least two, so called, "pivotal trials" in the same patient popula-

* Corresponding author: jpliu@email.stat.ncku.edu.tw
The views expressed in this article are professional opinions of the authors and may not necessarily represent the position of National Cheng-kung University, the National Health Research Institutes, National Cheng-Chi University, Taiwan and the US Food and Drug Administration, U.S.A.

 0323-3847/02/0812-0969 $ 17.50+.50/0

tion are evaluated as recommended by the U.S. Food and Drug administration (FDA). After a pharmaceutical product is approved by the regulatory agency in the original region, such as the United States or European Union, the sponsor might seek registration of the product in a new region, e.g., an Asian country. Because of different ethnicity and clinical practice in the new region, necessity of repeating all or any of phase I, phase II and phase III clinical trials with the same scale in the new region has been discussed and debated. Recently the International Conference on Harmonisation (ICH) has published a tripartite guidance entitled "*Ethnic Factors in the Acceptability of Foreign Clinical Data*" to address the above issues (ICH E5, 1997).

The objective of the guidance is to provide a framework for evaluation of the impact of ethnic factors on the efficacy and safety of a pharmaceutical product at a particular dosage or dose regimen. The guidance describes regulatory strategies to minimize duplication of clinical data and the requirement of bridging evidences to allow extrapolation of the foreign clinical data to the population of the new region. According to the ICH E5 guidance, a bridge data package consists of (a) selected information from the Complete Clinical Data Package (CCDP) from the foreign region that is relevant to the population of the new region, and (b) if needed, a bridging study designed to extrapolate the foreign efficacy and/or safety data to the new region. In other words, bridging data can be obtained by two strategies: they are either in the CCDP generated during clinical drug development for submission to the original region or from the bridging studies conducted in the new region after the pharmaceutical product was approved in the original region.

In addition, the ICH E5 provides general guidance about the ability to extrapolate data generated from a bridging study: (i) If the bridging study shows that dose response, safety and efficacy in the new region are similar, then the study is readily interpreted as capable of "bridging" the foreign data. (ii) If a bridging study, properly executed, indicates that a different dose in the new region results in a safety and efficacy profile that is not substantially different from that derived in the original region, it will often be possible to extrapolate the foreign data to the new region, with appropriate dose adjustment, if this can be adequately justified. As a result, the ability of extrapolation of the foreign data to the new region depends upon similarity between the new and original regions. Although the ICH E5 guidance does not provide a precise definition of similarity, it does require that the safety and efficacy profile of the new region be not substantially different from that of the original region. Similarity is therefore interpreted in ICH E5 as "no substantial difference" which can be statistically interpreted as equivalence. In addition, recently the equivalence testing received much attention to evaluate therapeutic equivalence (Durrleman and Simon, 1990; Dunnett and Gent, 1977; Blackwell, 1982; Liu, 1995; Jennison and Turnbull, 1993; Fleming, 2000; Rohmel, 1998, Ebbutt and Frith, 1998; and Siegel, 2000). As a result, the concept of two-sided equivalence or average bioequivalence (Chow and Liu, 2000, and ICH E9, 1998) is appropriate to evaluate the similarity required by the ICH

E5 guidance. In what follows, we use the term "similarity" and "equivalence" interchangeably.

Under the strategy (a), a randomized parallel dose-response design with region as a stratified factor can be employed to prospectively and concurrently generate information on dose response, efficacy and safety during the clinical drug development stage. This strategy will provide the most convincing bridging evidence that is internally valid for evaluation of the similarity between the regions with respect to dose response, efficacy and safety. If the similarity between the regions can be verified by the data provided by this strategy, then there is no need to conduct a bridging study in a new region. However, the sample size required under this strategy, as shown later, will be much larger than that for the assessment of the overall dose response, efficacy and safety.

When the data provided in the CCDP are not adequate for extrapolation or the pharmaceutical product is ethnically sensitive, then bridging studies should be conducted in the new region to generate data needed to bridge the clinical data between the two regions. A bridging study could consist of another efficacy trial conducted in the new region. Under the strategy (b), a bridging data set would comprise clinical data generated from the new region as well as those by the clinical trials from the original region. The goal of the analysis, therefore, is not only to provide information on the dose response, efficacy, and safety with the data from the new region but also to evaluate similarity of dose response, efficacy and safety with the foreign clinical data conducted in the original region before the approval.

Since the main objective of a bridging strategy is to minimize the duplication of clinical data generated in the new region, the sample size becomes a very important issue. To reduce the sample size of a bridging study, information on the dose response, efficacy and safety from the original region can and should be incorporated in a statistically sound manner to evaluate bridging evidence. CHOW, et al. (2001) suggested using a reproducibility probability to evaluate bridging evidence. The reproducibility probability provides the power of detecting a treatment difference with an adjustment over possible range of the difference and variability expected in the new region. However, this approach fails to address the similarity between the new and original regions with respect to dose response, safety and efficacy as required by the ICH E5 guidance.

The purpose of this paper is to address the issues on sample size in planning bridging study and evaluation of bridging evidence. In Section 2, we formulate the hypothesis to assess the ability to extrapolate bridging evidence from the original region to the new region in terms of testing for similarity. In addition, the issue on the sample size for assessment of the similarity based on the data prospectively generated in the same trial during the clinical drug development stage (strategy a) is addressed. In Section 3, we present a hierarchical model to evaluate the sample size required for the analysis of bridging study (strategy b). Section 4 provides a numerical example for illustration. Discussion and final remarks are given in the last section.

2. Extrapolation and Similarity

For simplicity, here we only consider the problem for an assessment of efficacy in a comparison between a test treatment with a placebo control. However, the methods derived here can be directly applied to evaluate similarity of dose response or safety between the new and original regions. Suppose that a global pharmaceutical company currently develops a new drug for treatment of patients with chronic hepatitis B. A randomized double-blind study is being conducted in both the northern America region and the Asian Pacific region to compare the new drug treatment with a placebo as the concurrent control. One of the primary efficacy endpoints for the degree of hepatic inflammation and fibrosis is the Knodell Histologic Activity Index (KHAI) based on the liver biopsy (KNODELL, 1981). The KHAI has a range of 0 to 18 with higher scores indicating more severe abnormality.

Let N be the total number of patients in the study with N_O patients recruited from the northern America region and N_N patients from the Asian Pacific region, where N_O and N_N are assumed to be even, and $N_O + N_N = N$. Let Y_{ijk} be the KHAI for patient k receiving treatment j in region i, $k = 1, \ldots, K$, $j = T$ (test), P (placebo), and $i = O$ (northern America), N (Asian Pacific); $K = N_O/2$ if $i = O$ and $K = N_N/2$ if $i = N$. We assume that Y_{ijk} is independent and normally distributed with mean μ_{ij} and variance σ^2. The treatment effect (difference) at region i is defined as

$$\Delta_i = \mu_{iT} - \mu_{iP}, \qquad i = O, N. \tag{1}$$

Following FLEISS (1986) and YATES (1934), the overall treatment effect is defined as the simple average over the treatment effects of the two regions:

$$\Delta = (1/2)\,[(\mu_{OT} - \mu_{OP}) + (\mu_{NT} - \mu_{NP})] \tag{2}$$

Under the assumption of no region-by-treatment interaction, it gives $\Delta = (\mu_{OT} - \mu_{OP}) = (\mu_{NT} - \mu_{NP}) = (\mu_T - \mu_P)$. The hypothesis of testing for an overall treatment is given as

$$H_O: \mu_T - \mu_P = 0 \quad \text{vs.} \quad H_a: \mu_T - \mu_P \neq 0 \tag{3}$$

The statistic $d = (1/2)\,[(Y_{OT} - Y_{OP}) + (Y_{NT} - Y_{NP})]$ has the mean Δ and variance $\sigma^2(1/N_O + 1/N_N)$, where Y_{ij} is the sample mean for the treatment j in region i. A usual test statistic for hypothesis (3) is then given as $T_d = d/(\sigma\sqrt{1/N_O + 1/N_N})$. The null hypothesis in (3) is rejected if $|T_d| > z(\alpha/2)$, where $z(\alpha/2)$ is the $\alpha/2$-th upper percentile of the standard normal distribution. The total sample size required to achieve the $(1 - \beta)$ power of detecting the treatment difference Δ at the two-sided α level test is

$$N_d \geq \sigma^2/[\Delta^2 r(1 - r)]\,[z(\alpha/2) + z(\beta)]^2, \tag{4}$$

where $r = N_O/N$ and $(1 - r) = N_N/N$. Note that the statistic d is always valid for testing the overall treatment effect regardless whether or not a region-by-treatment interaction exists.

The similarity with respect to the efficacy can be interpreted as the difference of treatment effects between the two regions is within some clinically acceptable limit, say δ. The relationship between δ and Δ can be expressed as $\delta = f\Delta$. δ is clinically acceptable and meaningful only if $0 < f < 1$ because similarity dictates that the difference of treatment effects between regions should be smaller than the overall treatment effect. Let $\theta = (\mu_{NT} - \mu_{NP}) - (\mu_{OT} - \mu_{OP})$ denote the difference of treatment effects between the two regions. A hypothesis for evaluation of the similarity between the two regions can be formulated as the two-sided equivalence hypothesis:

$$H_O : \theta \leq -\delta \quad \text{or} \quad \theta \geq \delta \quad \text{vs.} \quad H_a : -\delta < \theta < \delta . \tag{5}$$

Denote the sample estimate of θ as $t = (Y_{NT} - Y_{NP}) - (Y_{OT} - Y_{OP})$. The sample estimate t has a normal distribution with the mean θ and variance $v(t) = 4\sigma^2/[Nr(1 - r)]$. Define two test statistics

$$T_L = (t + \delta)/\sqrt{v(t)}, \quad \text{and} \quad T_u = (t - \delta)/\sqrt{v(t)} \tag{6}$$

The null hypothesis (5) is rejected and similarity between the new and original region is concluded at the α significance level if and only if $T_L > z(\alpha)$ and $T_U < -z(\alpha)$. When $\theta = 0$, the total sample size required to achieve $(1 - \beta)$ power for testing the similarity hypothesis (5) is given as

$$\begin{aligned} N_s &\geq \{4\sigma^2/[\delta^2 r(1 - r)]\}\, [z(\alpha) + z(\beta/2)]^2 \\ &= \{4\sigma^2/[f^2\Delta^2 r(1 - r)]\}\, [z(\alpha) + z(\beta/2)]^2 . \end{aligned} \tag{7}$$

If Δ represents the overall treatment effect, the ratio of the total sample size required for testing similarity (5) to that for testing an overall treatment effect (3) is given as

$$N_s/N_d = (4/f^2)\, \{[z(\alpha) + z(\beta/2)]^2/[z(\alpha/2) + z(\beta)]^2\} \tag{8}$$

When $\alpha = 0.05$ and $\beta = 0.20$, the ratio of $[z(\alpha) + z(\beta/2)]^2/[z(\alpha/2) + z(\beta)]^2$ is about 1.09. Consequently, the ratio of the two sample sizes is approximately

$$N_s/N_d \cong 1.09(4/f^2) .$$

Two remarks can be made from the above equation. First, the ratio is independent of the stratified fraction r. Second, if f is chosen as 0.5, a rather liberal equivalence limit, then the total sample size required for evaluation of the similarity is at least sixteen times as large as that for the assessment of overall treatment effect. The test for similarity requires a much larger sample size that represents a tremen-

dous burden in resources for clinical drug development. Unless the market is extremely important to the sponsor, the strategy of conducting a local bridging study after the approval of the drug product in the original region might represent a cost-effective alternative.

3. Hierarchical Models

Suppose that I trials have been conducted for the approval of the drug product in the original region and a bridging trial is being conducted in a new region for registration after the approval in the original region. Let Y_{ijk} be the clinical response for patient k receiving treatment j on the ith trial conducted in the original region, $k = 1, \ldots, n_{ij}$, $j =$ T (test), P (placebo), and $i = 1, \ldots, I$. Assume that Y_{ijk}'s are independently normally distributed with mean μ_{ij} and variance σ_{ij}^2, $k = 1, \ldots, n_{ij}$, $j =$ T (test), P (placebo), and $i = 1, \ldots, I$. Following TARONE (1982) and PRENTICE, et al. (1992), we further assume that μ_{ij} has a normal distribution with mean μ_{Oj} and variance ν_{Oj}^2, $j =$ T, P. Consequently, the Y_{ijk}'s are independently normally distributed with mean μ_{OJ} and variance $\omega_{ij}^2 = \sigma_{ij}^2 + \nu_{Oj}^2$, $J =$ T, P. Similarly, let Y_{Njk} be the clinical response from a bridging study conducted in the new region, $k = 1, \ldots, n_{Nj}$, $j =$ T (test), P (placebo). Again, Y_{Njk}'s are assumed to be independently normally distributed with mean μ_{Nj} and variance ω_{Nj}^2, $J =$ T, P.

Define Y_{ij} be the sample mean for treatment j in trial i, $j =$ T, P; $i = 1, \ldots, I$. The maximum likelihood estimate (MLE) of μ_{OJ} is given as

$$t_{Oj} = [\Sigma Y_{ij}/(w_{ij}^2/n_{ij})]/\{\Sigma[1/(w_{ij}^2/n_{ij})]\}, \tag{9}$$

where $w_{ij}^2 = \Sigma(Y_{ijk} - t_{Oj})^2/n_{ij}$ is the MLE of ω_{ij}^2, $j =$ T, P; $i = 1, \ldots, I$. The MLEs t_{Oj} and w_{ij}^2 can be solved iteratively. For the bridging study, the MLE of the mean μ_{Nj} is the sample mean Y_{Nj}, denoted by t_{Nj}, $j =$ T, P. The MLEs t_{Oj} and t_{Nj} are independent and follow normal distributions with asymptotic variances $v(t_{Oj}) = 1/\{\Sigma[1/(\omega_{ij}^2/n_{ij})]\}$ and $v(t_{Nj}) = \omega_{Nj}^2/n_{Nj}$. The estimated asymptotic variances for t_{Oj} and t_{Nj} are $s_{Oj}^2 = 1/\{\Sigma[1/(w_{ij}^2/n_{ij})]\}$ and $s_{Nj}^2 = \Sigma(Y_{Nik} - t_{Nj})^2/n_{Nj}^2$, respectively. Thus, the statistic $t = (t_{NT} - t_{NP}) - (t_{OT} - t_{OP})$ is an asymptotic unbiased estimate for θ with the estimated asymptotic variance $s^2 = s_{NT}^2 + s_{NP}^2 + s_{OT}^2 + s_{OP}^2$. The test statistics for similarity (5) under the hierarchical model are given by

$$T_L = (t + \delta)/s \quad \text{and} \quad T_U = (t - \delta)/s\,.$$

The null hypothesis is rejected and similarity between the new and original regions is concluded if and only if $T_L > z(\alpha)$ and $T_U < -z(\alpha)$. Alternatively, the $(1 - 2\alpha)\%$ confidence interval for θ given as (L, U), where $U = t + z(\alpha)\,s$ and $L = t - z(\alpha)\,s$. The similarity between the new and original region is claimed at α significance level if and only if (L, U) is completely contained within $(-\delta, \delta)$.

Note that the test statistics T_L and T_U involve an external cross-trial comparison. They are based on the constancy assumption that the treatment effect is unchanged if the same patient population were enrolled in the bridging study. In addition, the factors other than ethnic factors may contribute to the variability between the studies. Therefore, to minimize bias, it is critical to keep the design characteristics and conduct of the bridging study as close to the studies previously conducted in the original studies as possible.

If n_N is the total number of patients in the bridging study with n_{Nt} patients in the treatment group and N_{Np} in the control group, denote the allocation fraction for the treatment as $g_{NT} = n_{Nt}/n_N$. Let

$$A_1 = s^2_{NT}/g_{NT} + s^2_{NP}/(1 - g_{NT}),$$
$$A_2 = \delta^2/[z(\alpha) + z(\beta/2)]^2 \tag{10}$$

and

$$A_3 = s^2_{OT} + s^2_{OP}.$$

It can be shown that when $\theta = 0$, the total sample size required for a bridging study to achieve $(1 - \beta)$ power for testing the similarity hypothesis (5) is given as

$$n_N \geq A_1/(A_2 - A_3) \tag{11}$$

If $w^2_{ij} = w^2_{Nj} = w^2$, for all i and j, $\Sigma n_{iT} = \Sigma n_{iP}$, denoting that $CV = 2w/(\mu_{OT} - \mu_{OP})$ and $\delta = f(\mu_{OT} - \mu_{OP})$, then the formula for sample size in (11) can be simplified as

$$n_N = [1/g_{NT}(1 - g_{NT})]/\{(2f/CV)^2/[z(\alpha) + z(\beta/2)]^2 - (4/N_O)\}, \tag{12}$$

where $0 < f < 1$ and $N_O = \Sigma n_{iT} + \Sigma n_{iP}$ is the total number of patients enrolled in the trial conducted in the original region,

Table 1 presents the total sample size for a bridging study computed according to formula (12) for $\alpha = 0.05$, $\beta = 0.2$ and $g_{NT} = 1/2$ for various combinations of CV, N, and f. Several remarks can be made from the equations (11) and (12) and Table 1. First because the $g_{NT}(1 - g_{NT})$ in (12) is maximized when $g_{NT} = 1/2$, it follows that the total sample size for the bridging study in the new region is minimized if an equal allocation is employed. When $g_{NT} = 1/2$,

$$n_N \geq 1/\{(f/CV)^2/[z(\alpha) + z(\beta/2)]^2 - (1/N_O)\} \tag{13}$$

Second, as the CV increases, the total sample size for the bridging study also increases. Finally, the denominator is an increasing function of the total number of patients enrolled in the trials conducted in the original region. As a result, if the total number of patients N_O is sufficiently large, then the contribution of $1/N_O$ in the denominator is negligible. The total sample size may then be approximated by

$$n_N \geq (CV/f)^2 [z(\alpha) + z(\beta/2)]^2 \tag{14}$$

Table 1
Total sample size required for the two-sided equivalence test for the bridging study in the new region for $\alpha = 0.05$, $\beta = 0.2$ and $g_{NT} = 0.5$ by CV, N_O and f

CV	N_O	f				
		0.1	0.2	0.3	0.4	0.5
40%	400	208	38	16	10	6
	1000	160	36	16	10	6
	3000	144	36	16	10	6
	5000	142	34	16	10	6
80%	400	–	208	72	38	24
	1000	1214	160	66	36	22
	3000	672	144	62	36	22
	5000	604	142	62	34	22
100%	400	–	462	126	62	38
	1000	5964	272	106	58	36
	3000	1200	232	100	54	36
	5000	1034	224	98	54	36
200%	400	–	–	7854	462	208
	1000	–	5964	616	272	160
	3000	–	1090	436	232	144
	5000	10878	1000	412	224	142
300%	400	–	–	–	–	1344
	1000	–	–	5964	930	446
	3000	–	5388	1200	574	344
	5000	–	3136	1034	534	330

$CV = 2w/(\mu_{OT} - \mu_{OP})$, w is the common variance of the response. N_O is the total number of patients enrolled in the trials with equal allocation in the trials conducted in the original region. f is the proportion of the clinically acceptance limit δ expressed as a function of the mean difference between the test and placebo i.e. $\delta = f(\mu_{OT} - \mu_{OP})$.

On the other hand, the non-inferiority test may also be appropriate for evaluating similarity between the two regions because the efficacy of new region can be claimed similar if it is not inferior to that of the original region. The one-sided non-inferiority hypothesis is given as

$$H_{OL} : \theta \leq -\delta \quad \text{vs.} \quad H_a : \theta > -\delta \tag{15}$$

The non-inferiority of the new region can be claimed at the α significance level if $T_L > z(\alpha)$. The sample size for the non-inferiority can be obtained by substituting $\beta/2$ by β in formula (12). Table 2 provides the total sample size for the non-inferiority test for a bridging study for $\alpha = 0.05$, $\beta = 0.2$ and $g_{NT} = 1/2$ for various combinations of CV, N, and f. From Table 2, the sample sizes required for non-inferiority test are in general one third fewer than those for the two-sided equivalence test in evaluation of bridging evidence for similarity between the new and original region.

Table 2

Total sample size required for non-inferiority test for the bridging study in the new region for $\alpha = 0.05$, $\beta = 0.2$ and $g_{NT} = 0.5$ by *CV*, N and *f*

CV	N_O	*f*				
		0.1	0.2	0.3	0.4	0.5
40%	400	132	26	12	6	4
	1000	110	26	12	6	4
	3000	102	26	12	6	4
	5000	102	26	12	6	4
80%	400	36668	132	50	26	16
	1000	656	110	46	26	16
	3000	456	102	46	26	16
	5000	430	102	46	26	16
100%	400	–	252	84	44	26
	1000	1620	184	74	40	26
	3000	780	164	70	40	26
	5000	706	160	70	40	26
200%	400	–	–	878	252	132
	1000	–	1620	380	184	110
	3000	14080	780	302	164	102
	5000	4894	706	292	160	102
300%	400	–	–	–	2664	502
	1000	–	–	1620	534	288
	3000	–	2594	780	394	240
	5000	–	1928	706	374	234

$CV = 2w/(\mu_{OT} - \mu_{OP})$, w is the common variance of the response. N_O is the total number of patients enrolled in the trials with equal allocation in the trials conducted in the original region. f is the proportion of the clinically acceptance limit δ expressed as a function of the mean difference between the test and placebo i.e. $\delta = f(\mu_{OT} - \mu_{OP})$.

4. Numerical Examples

A drug sponsor plans to conduct a randomized double-blind trial to investigate similarity on efficacy of a new drug versus the placebo in the patients with mild to moderate essential hypertension between the northern America and Asian Pacific regions. Based on sitting diastolic blood pressure (SDBP), it expects that after 24 weeks of treatment, the mean SDBP for the new drug will decrease from baseline by an amount of 15 mm/Hg while the mean SDBP in the placebo will decrease by 4. As a result, the relative efficacy of the new drug as compared to placebo is an improvement of mean reduction of SDBP from baseline of 11 mm/Hg. The standard deviation of change from baseline in SDBP is around 11. This gives a *CV* of 200.0%. The Northern American and Asian Pacific regions can be claimed as having a similar efficacy if the difference of the relative efficacy between the two regions is -5.5 mm/Hg, which is 50% of 11 mm/Hg. If an equal

number of patients is planned to be recruited from both regions, according to (7), a total of 548 patients would be required to achieve the 80% power for testing the similarity (5) at the 5% significance level.

Suppose the sponsor has conducted three randomized placebo controlled trials in the northern American region with a total of 918 patients to establish efficacy and safety of the new antihypertensive drug. The results in reduction of SDBP from baseline of the three trials of a particular dose are summarized in Table 3. Under the hierarchical model, μ_{OT} is estimated -16.9 mm/Hg and μ_{OP} is -3mm/Hg. It follows that the estimated relative efficacy is -13.9 mm/Hg with an estimated asymptotic variance of 0.58. This gives a value of -18.2 for the z statistic with a p-value < 0.0001 which clearly establishes the efficacy of the drug in reduction of SDBP. After the drug was approved by the Northern America region, a bridging study in the Asian Pacific region is planned. Using equation (11), a total sample size of 148 for the bridging is required to achieve 80% power for evaluation of similarity between the two regions with the same equivalence limit of 5.5 mm/Hg at the 5% significance level.

For illustration purpose, the bottom of Table 3 presents the results of a bridging study conducted in an Asian Pacific country with the same study design, the same inclusion/exclusion criteria, the same dose and dosing regimen, and the same primary endpoints. Since the primary endpoint is the reduction from baseline in SDBP, if the same equivalence limit of 5.5 mm Hg is employed, the non-inferiority hypothesis becomes $H_{\mathrm{OL}}: \theta \geq 5.5$ vs. $H_{aL}: \theta < 5.5$. It can be easily verified that $T_L = 3.59 > -1.645$. We conclude that the mean reduction from baseline in SDBP in the new region can not be concluded to be non-inferior to that of the original region at the 5% significance level. Consequently the relative efficacy as measured by the mean reduction from baseline in SDBP is not similar to that of the original region.

Table 3
Summary of Reduction from Baseline in Sitting Diastolic Blood Pressure

Region	Study	Statistics	Treatment Group	
			Drug	Placebo
Original	1	N	138	132
		Mean (mm Hg)	−18	− 3
		Standard Deviation	11	12
	2	N	185	179
		Mean (mm Hg)	−17	− 2
		Standard Deviation	10	11
	3	N	141	143
		Mean (mm Hg)	−15	− 5
		Standard Deviation	13	14
New	1	N	64	65
		Mean (mm Hg)	− 4.7	− 3.8
		Standard Deviation	11	11

5. Discussion

The ICH E5 guidance provides regulatory requirements that the acceptability of foreign data should be evaluated based on the similarity of dose response, efficacy and safety between the foreign and new regions. A scientifically sound and statistically valid interpretation of similarity is the hypothesis for the two-sided equivalence in (5). The two one-sided tests (TOST) procedure is presented for evaluation of similarity between the regions concurrently in a prospectively randomized trial. Under the hierarchical model, TOST is also suggested for evaluation of equivalence between the results of the bridging study conducted in the new region and those from the original region. The formulas for sample size determination are also derived. If a sponsor chooses to evaluate the similarity between two regions concurrently in a prospectively randomized trial, the sample size required can be quite formidable. Consequently, until some innovative Bayesian procedures are proposed, the traditional TOST for evaluation of bridging evidence generated from a prospectively randomized trial may be of little practical value due to the prohibitively large sample size.

The similarity requirement by the ICH E5 guidance is to evaluate whether across-trial difference of treatment effect relative to placebo (test-placebo) is within some pre-specified clinically meaningful limit. There are lots of publications on the controversial selection of so called clinically meaningful equivalence limits for active controlled "non-inferiority" trials, e.g., TEMPLE and ELLENBERG (2000), ELLENBERG and TEMPLE (2000), FISHER et al. (2001), HUNG (2001), SNAPPIN (2001), TSONG, et al. (2001), WITTES (2001). However, if the equivalence limit is based on the point estimate of the difference between the test product and placebo from the results of the clinical trial conducted in the original region, then its variability should be taken into account. Otherwise, the type I error rate will be inflated. This phenomenon was noted in LIU and WENG (1995). The use of the lower limit of the confidence interval for the difference between the test product and placebo is one way to alleviate this problem. LIU and WENG (1995) suggested the following alternative approach. If $\delta = f(\mu_{OT} - \mu_{OP})$, then the two-sided equivalence hypothesis in (5) can be decomposed into the following two one-sided hypotheses:

$$\begin{aligned} &H_{LO}: (\mu_{NT} - \mu_{NP}) - (1-f)(\mu_{OT} - \mu_{OP}) \leq 0 \quad \text{vs.} \\ &H_{LA}: (\mu_{NT} - \mu_{NP}) - (1-f)(\mu_{OT} - \mu_{OP}) > 0 \end{aligned}$$

and (16)

$$\begin{aligned} &H_{UO}: (\mu_{NT} - \mu_{NP}) - (1+f)(\mu_{OT} - \mu_{OP}) \geq 0 \quad \text{vs.} \\ &H_{UA}: (\mu_{NT} - \mu_{NP}) - (1+f)(\mu_{OT} - \mu_{OP}) < 0\,. \end{aligned}$$

Under the hierarchical model, it is then straightforward to obtain the test statistics that are given as

$$T'_L = t_L/s_L \quad \text{and} \quad T'_U = t_U/s_U\,,$$

where $t_L = (t_{NT} - t_{NP}) - (1-f)(t_{OT} - t_{OP})$, $t_U = (t_{NT} - t_{NP}) - (1+f)(t_{OT} - t_{OP})$, $s_L^2 = (s_{NT}^2 + s_{NP}^2) + (1-f)^2 (s_{OT}^2 + s_{OP}^2)$ and $s_L^2 = (s_{NT}^2 + s_{NP}^2) + (1+f)^2 \times (s_{OT}^2 + s_{OP}^2)$.

The null hypothesis is rejected and similarity between the new and original regions is concluded if and only if $T_L > z(\alpha)$ and $T_U < -z(\alpha)$. The power function for the above procedure based on T_L' and T_U' is quite complicated and depends on the nuisance parameter $(\mu_{OT} - \mu_{OP})$. The performance of size and power and sample size determination for this procedure requires further research analytically and empirically.

As shown in formula (12), based on the hierarchical model, the sample size required for the bridging study in the new region is a function of f, CV, and N_O. In addition, if when $g_{NT} = 1/2$ and $(f/CV)^2/[z(\alpha) + z(\beta/2)]^2 \leq (1/N_O)$, then n_N is negative. In other words, no sample size for the bridging study in the new region can achieve $1-\beta$ power for testing hypothesis (5) at the α significance level. This phenomenon will occur when the variability of the endpoint for evaluation of efficacy (i.e., CV) is too large; the equivalence limit is too strict, and the size of foreign clinical database in the original region is small. For example, from Table 1, no sample size exists for when $f = 0.1$ or 0.2; CV is greater than 200%; and N_O is fewer than 1000. Furthermore, although the sample size can be obtained when CV is greater than 200% for $f > 0.3$, they usually are quite large and exceed 1000. As a result, the bridging study in the new region might be feasible if the CV of the primary efficacy endpoint is less 200% and the number of patients is at least 1000 in the foreign clinical data from the original region.

Acknowledgements

We want to thank the Editor and two anonymous referees for their constructive comments and suggestions.

References

BLACKWELDER, W. C., 1982: "Prove the null hypothesis" in clinical trials. *Controlled Clinical Trials*, **3**, 345–353.

DURRLEMAN, S., and SIMON. R., 1990: Planning and Monitoring of equivalence studies. *Biometrics*, **46**, 329–336.

DUNNETT, C. W., and GENT, M., 1977: Significance testing to establish equivalence between treatments with special reference to data in the form of 2×2 table. *Biometrics*, **33**, 593–602.

CHOW, S. C., and LIU, J. P., 2000: *Design and Analysis of Bioavailability and Bioequivalence Studies*, 2nd Edition, Revised and Expanded, Marcel Dekker, Inc., New York, New York.

CHOW, S. C., SHAO, J., and HU, O. Y. P., 2001: Bridging Studies for Clinical Development, Submitted for publication.

EBBUTT, A. F., and FRITH, L., 1998: Practical issue in equivalence trials. *Statistics in Medicine*, **17**, 1691–1701.

ELLENBURG, S. S. and TEMPLE, R., 2000: Placebo-controlled trials and active-controlled trials in the evaluation of new treatment, part 2: practical issue and specific cases. *Annals of Internal Medicine* **133**, 464–470.

FISHER, L. D., GENT, M., and BULLER, H. R., 2001: Active-control trials: how would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin, and placebo. *American Heart Journal* **141**, 26–32.

FLEISS, J. L., 1986: Analysis of data from multiclinic trials, *Controlled Clinical Trials* **7**, 267–275.

FLEMING, T. R., 2000: Design and interpretation of equivalence trials. *American Heart Journal* **139**, 172–176.

ICH, 1997: International Conference on Harmonisation Tripartite Guidance E 5 on *Ethnic Factors in the Acceptability of Foreign Data*.

ICH, 1998: International Conference on Harmonisation Tripartite Guidance E 9 on *Statistical Principles for Clinical Trials.*

HUNG, H. M. J., 2001: Noninferiority: A dangerous toy. *ICSA Bulletin,* Jan., 27–29.

JENNISON, C., and TURNBULL, B. W., 1993: Sequential equivalence testing and repeated confidence intervals, with applications to normal and binary responses. *Biometrics* **49**, 31–43.

KNODELL, R. G., ISHAK, K. G., BLACK, W. C., et al., 1981: Formulation and application of a numerical scoring system for assessing histological activity in asymptotic chronic active hepatitis, *Hepatology* **1**, 431–435.

LIU, J. P., 1995: Letter to the Editor on "Sample size for therapeutic equivalence based on confidence interval" by S.C. Lin. *Drug Information Journal* **29**, 1063–1064.

LIU, J. P. and CHOW, S. C., 1992: Sample size determination for the two one-sided tests procedure in bioequivalence, J*ournal of Pharmacokinetics and Biopharmaceutics* **20**, 101–104.

LIU, J. P. and WENG, C. S., 1995: Bias of two one-sided tests procedures in assessment of bioequivalence. *Statistics in Medicine* **14**, 853–862.

MAKUCH, R. W. and SIMON, R., 1978: Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Report* **6**, 1037–1040.

MAKUCH, R. W. and JOHNSON, M., 1989: Issues in planning and interpreting active control equivalence studies. *Journal of Clinical Epidemiology* **42**, 503–511.

PRENTICE, R. L., SMYTHE, R. T., KREWSKI, D., and MASON, M., 1992: On the use of historical control to estimate dose response trends in quantal bioassay, *Biometrics* **48**, 459–478.

ROHMEL, J., 1998: Therapeutic equivalence investigations: statistical considerations. *Statistics in Medicine*, **17**, 1703–1714.

SIEGEL, J. P., 2000: Equivalence and noninferiority. *American Heart Journal* **139**, 166–170.

SIMON, R., 1999: Bayesian design and analysis of active control clinical trials, *Biometrics* **55**, 484–487.

SNAPINN, S. M., 2001: Alternative for discounting historical data in the analysis of noninferiority trial. *ICSA Bulletin* Jan.**,** 29–33.

TARONE, R. E., 1982: The use of historical control information in testing for a trend in proportions, *Biometrics* **38**, 215–220.

TEMPLE, R. and ELLENBURG, S. S., 2000: Placebo-controlled trials and active-controlled trials in the evaluation of new treatment, part 1: ethnical and scientific issues. *Annals of Internal Medicine* **133,** 464–470.

TSONG, Y., WANG, S. J., CUI, L., and HUNG, J. H. M., 2001: Placebo control, historical control, and active control trials, *ICSA Bulletin* Jan.**,** 36–39.

WITTES, J., 2001: Active-control trials: a linguistic problem. *ICSA Bulletin* Jan.**,** 39–40.

YATES, F., 1934: The analysis of multiple classifications with unequal numbers in the different classes, *Journal of the American Statistical Association* **29**, 51–66.

Received, March 2001
Revised, February 2002
Accepted, July 2002