Impact of the ICH E9 Guideline *Statistical Principles for Clinical Trials* on the Conduct of Clinical Trials in Japan

Chihiro Hirotsu, PhD
Meisei University, Tokyo, Japan

Ludwig A. Hothorn, PhD
University of Hannover, Hannover, Germany

This article evaluates the impact of the ICH E9 guideline Statistical Principles for Clinical Trials *on the conduct of clinical trials in Japan. In particular, the following Japanese practices in the conduct of clinical trials are discussed in detail from the ethical, statistical, and logical viewpoints: 1. Conducting only one phase 3 multicenter trial with many centers and few subjects per center; 2. Seeking to show noninferiority to an active control rather than superiority to placebo; and 3. Choosing a global assessment variable with a subjective component as the primary endpoint. The influence of public health insurance and the potential number of patients in Japan on various aspects of a trial are discussed. Problems requiring further research are mentioned and points requiring clarification are highlighted.*

Key Words
Noninferiority trial; Random and fixed effects models for centers; Subjective and objective measures

Correspondence Address
Prof. Dr. C. Hirotsu, Faculty of Science & Technology, Meisei University, 2-1-1 Hodokubo, Hino-City, Tokyo 191-8506, Japan (e-mail: hirotsu@ge.meisei-u.ac.jp).

INTRODUCTION

The ICH E9 guideline *Statistical Principles for Clinical Trials* (1) provides common principles for statistical methodology for clinical trials to harmonize drug development in Europe, the United States, and Japan. Recently, the British group Statisticians in the Pharmaceutical Industry discussed several consequences of this guideline, including consequences related to multicenter trials (2). Although the guideline takes a view that is broadly inclusive of Japanese practice, there are several practices in the conduct of clinical trials in Japan that differ from those in Europe and the United States. These practices include:

1. Conducting only one phase 3 multicenter trial with many centers and few subjects per center,
2. Seeking to show noninferiority to an active control rather than superiority to placebo, and
3. Choosing a global assessment variable with a subjective component as the primary endpoint.

We will provide typical examples of these past practices in Japan.

The first example is a phase 3 trial conducted in 1991 and 1992 aimed at proving the noninferiority of Trandolapril to Enalapril, the active control, in lowering blood pressure. This was a randomized, double-blind trial with 299 patients from 118 research institutes throughout Japan. Four patients (two for each treatment), were randomly assigned to the two treatments in each institute by two controllers, who handled the random and double-blind allocation of cases. Upon the request of the pharmaceutical sponsor, a central committee composed of 13 representatives of those institutes, including the two controllers, supervised all aspects of the trial. Their work included protocol adherence and deciding how to treat abnormal cases, such as violations of the protocol. A central office that was independent of the pharmaceutical sponsor managed the clerical work and controlled the data collected.

The central committee chose the proportion of the blood pressure lowering effect category, based upon clinical judgment, as defined in Table 1, as the primary endpoint. In particular, a patient whose blood pressure matched the slightly lowered category was assigned to the lowered category if his blood pressure decreased below 150 mm Hg (systolic) and 90 mm Hg (diastolic). The confidence intervals for the difference between the two proportions in the lowered category with the two-sided confidence coefficient 0.90 were $-0.115 \leq \cdot \leq 0.07$ by intent-to-treat analysis and $-0.051 \leq \cdot \leq 0.131$ by the protocol compatible analysis; it was argued that these figures cleared the maximum tolera-

TABLE 1

Ordered Categories of Blood Pressure Lowering				
	Lowered	Slightly Lowered	Unchanged	Raised
Systolic blood pressure (mm Hg)	$\cdot < -20$	$-20 \cdot \leq \cdot < -9$	$-9 \leq \cdot \leq 9$	$9 < \cdot$
Diastolic blood pressure (mm Hg)	$\cdot < -10$	$-10 \leq \cdot < -4$	$-4 \leq \cdot \leq 4$	$4 < \cdot$
Average blood pressure (mm Hg)	$\cdot < -13$	$-13 \leq \cdot < -6$	$-6 \leq \cdot \leq 6$	$6 < \cdot$

ble difference of 0.10, although this was not the case for the intent-to-treat analysis.

The committee chose the amount of the change in blood pressures and the proportion of normalized patients as defined by blood pressure that had been lowered below 150 mm Hg (systolic) and 90 mm Hg (diastolic) as the secondary endpoints. The overall usefulness was evaluated in four ordered categories by considering both efficacy and safety. Adverse events were reported and compared.

The second example is a phase 3 trial conducted in 1995 and 1996 that compared Fenofibrate to Clinofibrate, the active control, in improving rates of serum lipids as defined in Table 2. A global assessment rating, instead of quantitative raw data for cholesterol or triglycerides, was employed as a primary endpoint. Attending physicians used a global improvement rating to assess five ordered categories of improvement rate in serum lipids and subjective and objective symptoms. The overall usefulness was also evaluated in five ordered categories by considering both efficacy and safety.

The trial included 236 patients from 42 research institutes. Six patients (three for each treatment), were randomly assigned to the treatment or the control by the controller in each institute. Other aspects of the trial, such as the central committee and the central office, were similar to the first example.

Generally speaking, there are two more or less contradictory aspects in comparative clinical trials:

1. Providing scientific proof that a test treatment is superior or not inferior to the control in efficacy and safety, and
2. Generalizability of the results from a clinical trial to patients in the real world.

For the purpose of providing scientific proof that a test treatment is superior or not inferior to the control, it is more efficient to reduce various variations due to noise factors such as age and severity of disease, the type and scale of centers, the skill of doctors, and measurement errors. To ensure generalizability of the results, variations of the noise factors should be taken into account. Then an endpoint with larger measurement errors might be preferred if it is more closely related to the trial's clinical endpoint. Those two aspects are sometimes called explanatory and pragmatic (3), respectively. Some compromise is necessary between these two approaches.

It seems that the Japanese attitude has been more pragmatic than explanatory in choosing a wide range of centers in one trial rather than a

TABLE 2

Ordered Categories of Improvement Rates in Serum Lipids					
	Excellent	Moderate	Slightly Improved	Unchanged	Deteriorated
Total cholesterol (%)	$\cdot \leq -15$	$-15 < \cdot \leq -10$	$-10 \leq \cdot < -5$	$-5 < \cdot < 5$	$5 \leq \cdot$
Triglyceride (%)	$\cdot \leq -30$	$-30 < \cdot \leq -20$	$-20 < \cdot \leq -10$	$-10 < \cdot < 10$	$10 \leq \cdot$
HDL cholesterol (mg/dl)	$10 \leq \cdot$	$7 \leq \cdot < 10$	$4 \leq \cdot < 7$	$-4 < \cdot < 4$	$\cdot \leq -4$

few selected centers, aimed at finding the optimal dose for practice in the phase 2 trial rather than proving a dose-response beyond that dose, or employing subjective judgment by an attending physician rather than a quantitative measure as a primary endpoint. Assuming that there is only one phase 3 trial, as used to be the case in Japan, a trial involving many centers would have been necessary although some justification for the random effects model for centers is certainly necessary. Also, a clinical trial should be acceptable ethically as well as scientifically.

In this paper we will, therefore, address the ethical, statistical, and logical issues concerning conducting only one phase 3 multicenter trial with many centers and few subjects per center; seeking to show noninferiority to an active control rather than superiority to placebo; and choosing a global assessment variable with a subjective component as the primary endpoint. We will also discuss whether and how to change the conduct of clinical trials in Japan.

THE NUMBER OF CENTERS INVOLVED IN A TRIAL

Having many centers with few subjects per center has been recognized as one of the most prominent features of Japanese clinical trials. The ICH E9 guideline recommends having a considerable number of subjects per center in order to evaluate the interaction effects between the drug and the center. The Japanese Ministry of Public Health and Welfare, in a question and answer document about the new guideline, uses a yardstick of 10 subjects per arm per center. It recommends conducting clinical trials in a few selected centers. Then the observation of no interaction cannot ensure that the same thing will occur in a larger trial. It should be noted that the type of interaction will depend upon the type of centers involved in the trial and a small interaction effect by a few selected excellent centers might not actually reflect the situation. Naturally, by increasing the number of centers, more bad centers will be involved, which will impact the interaction effect. A more quantitative discussion concerning this point will be provided later.

Further, the guideline assumes basically a fixed effects model for each center and implies that the interpretation of the observed interaction and also the absence of interaction are within the centers involved in the trial. It should, therefore, be essential to conduct several phase 3 trials, changing the types of centers involved, in order to ensure the generalizability of the results. In Japan, however, only one trial has traditionally been conducted in phase 3; further discussions about the number and type of centers, rather than the number of subjects per center, will be necessary.

One possibility is to have many randomly selected centers with a relatively small number of subjects per center and to assume random effects for the center. The center is then regarded as a noise factor with random effects in Taguchi's parameter design (4). Interaction cannot be interpreted like a fixed effects model and the treatment effects should be proven beyond the institutional variations.

Japan has a long history of conducting such trials, although the randomness might not be strictly satisfied and it will be useful to have some idea of the amount of drug-center interaction effects in those trials. Gould (5) argued that the random effects model is a reasonable and convenient approximation even though it is true that centers are not randomly sampled from some plausible population of centers. We assume, therefore, a two-way analysis of variance model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},\ i = 1, \cdots, a;\quad j = 1, \cdots, b;\ k = 1, \cdots, n_{ij},$$

with treatment (drug) effect α_i, center effect β_j, drug-center interaction $(\alpha\beta)_{ij}$, and error ε_{ijk} representing both subject variations and measurement errors. We obtain the estimates of variance components σ^2_β, $\sigma^2_{\alpha\beta}$, σ^2_e by Hirotsu's method (6), where we purposely use the unweighted analysis to compare the effects of large and small centers equally. We denote the total number of subjects by n. The method has been applied to recent clinical trials in Japan. Table 3 shows the results;

TABLE 3

Variations in Clinical Assessment Among Centers

Disease	a	b	n	m	$\widehat{\sigma_\beta^2}/\widehat{\sigma_e^2}$	$m\widehat{\sigma_{\alpha\beta}^2}+\widehat{\sigma_e^2}/\widehat{\sigma_e^2}$	p	UC1	UC2
1. Hypertension	2	62	192	1.33	0.040	1.22	0.21	1.69	1.52
Systolic	2	62	192	1.33	0.016	1.23	0.21	1.69	1.52
Diastolic	2	62	192	1.33	0.188	1.00	0.50	1.38	1.29
2. Hypertension	2	48	171	1.52	0.037	1.26	0.18	1.79	1.52
Systolic	2	48	171	1.52	0.117	1.06	0.41	1.50	1.33
Diastolic	2	48	171	1.52	0.026	1.15	0.29	1.63	1.41
3. Antibiotics	2	9	54	1.78	0.045	0.47	0.87	1.12	1.07
4. Allergy	2	24	249	4.29	0.063	0.86	0.65	1.36	1.08
5. Allergy*	3	33	210	1.77	0.024	1.14	0.27	1.53	1.30
6. Cancer	2	35	177	1.62	0.024	1.30	0.16	1.91	1.56
7. Cerebrovascular	2	99	557	2.39	0.029	1.30	0.04	1.62	1.26
8. Cerebrovascular	2	36	211	1.85	0.029	1.01	0.47	1.46	1.25
9. Neurosis*	3	9	47	1.38	0.172	1.20	0.35	2.26	1.92
HAM-A	3	9	47	1.38	0	0.82	0.65	1.56	1.41
10. Neurosis*	3	28	156	1.56	0	1.47	0.06	2.06	1.68
HAM-A	3	29	162	1.57	0.012	1.01	0.48	1.40	1.26
11. Depression	2	36	137	1.53	0.230	0.62	0.94	0.93	0.95
HAM-D	2	36	137	1.53	0.288	0.62	0.94	0.93	0.95
12. Depression*	3	20	96	1.34	0	1.40	0.16	2.14	1.85
HAM-D	3	20	95	1.34	0.162	1.02	0.48	1.56	1.42
13. Depression	2	45	192	1.63	0.020	1.05	0.41	1.49	1.30
HAM-D	2	42	178	1.63	0.027	1.42	0.09	2.03	1.63
14. Schizophrenia	2	25	132	2.08	0	1.44	0.11	2.29	1.62
BPRS	2	25	133	2.09	0	1.58	0.07	2.51	1.72
15. Schizophrenia	2	32	141	1.55	0.033	0.91	0.60	1.38	1.24
BPRS	2	32	141	1.55	0.146	0.50	0.98	0.76	0.85
16. Schizophrenia	2	17	262	3.85	0	3.49	0.00	6.05	2.31
BPRS	2	17	262	3.85	0	1.44	0.12	2.50	1.39

*Phase 2 trials
HAM-A: Hamilton Anxiety Scale
HAM-D: Hamilton Depression Scale
BPRS: Brief Psychiatric Rating Scale
Cancer: Shrinkage measurement
UC1: Upper 90% confidence interval for $(m\sigma_{\alpha\beta}^2+\sigma_e^2)/\sigma_e^2$.
UC2: Upper 90% confidence interval for $(\sigma_{\alpha\beta}^2+\sigma_e^2)/\sigma_e^2$.
p: p value for testing the null hypothesis $H_0: \sigma_{\alpha\beta}^2=0$.

only centers with at least one subject from each treatment group are included in the analysis and the * shows the phase 2 trials. The data are basically the Final Global Improvement Rate (FGIR) unless otherwise stated and the seven ordered categories are given merely the scores 1, ··· , 7, but this particular quantification will not affect the result much. Table 3 shows the ratios of estimated variance components $\widehat{\sigma^2_\beta}$, $m\widehat{\sigma^2_{\alpha\beta}}+\widehat{\sigma^2_e}$ and $\widehat{\sigma^2_{\alpha\beta}}+\widehat{\sigma^2_e}$ to $\widehat{\sigma^2_e}$ as scale invariant measures as well as the p value in testing $H_0: \sigma^2_{\alpha\beta}=0$, where $\widehat{\sigma^2_e}$ is the usual unbiased estimate of the error variance and m is the harmonic mean of n_{ij}. The upper 90% confidence intervals for $(m\sigma^2_{\alpha\beta}+\sigma^2_e)/\sigma^2_e$ and $(\sigma^2_{\alpha\beta}+\sigma^2_e)/\sigma^2_e$ are also given in the table; the latter provides the estimate of increased variation due to interaction in the extreme case of $m=1$.

Except for the mental disorder, the estimates $(\widehat{\sigma^2_{\alpha\beta}}+\widehat{\sigma^2_e})/\widehat{\sigma^2_e}$ are at most 1.5, which suggests that if the population response ranges between ±10 then the increased range due to the interaction will at most be ±12. For the mental disorder, a significant result is sometimes observed for testing the null hypothesis H_0 and the range goes up to ±15 either for the FGIR or the rating scales, but it is still not too large. Also, since the ratios of $m\widehat{\sigma^2_{\alpha\beta}}+\widehat{\sigma^2_e}$ to $\widehat{\sigma^2_e}$ are around 1.5 except for one case at the bottom of Table 3, an approximately 40 ~ 50% increase in the sample size can supplement the loss of power in detecting treatment effects due to the interaction as compared to the trial with a few very homogeneous centers. For other cases, a 10 ~ 20% increase in sample size will suffice.

The amount of the increase in the sample size will be a factor in considering the generalizability of the results of a single trial. One should also refer to Gould (5) for comparisons between fixed and random effects models, as well as some meta-analytic approaches. The fact that the institution variation is rather smaller than the error variation due to subjects and measurements is one of the eminent characteristics of the clinical measurements as compared with the more exact physical measurements where institutional variation will dominate the sample variation.

Of all the examples in Table 3, only #16 has a considerable number (five) of centers with more than 10 subjects per arm. Thus, we performed subgroup analysis for those centers as well as for the rest of the centers in the trial. Table 4 shows the result. There is a remarkable difference in the FGIR between the two groups, with distinguishable interaction effects only among those centers with fewer subjects. Of course, the number of subjects recruited by one center might vary just by random fluctuation (7). In this example, however, the largest number of subjects per center was 39, whereas 5 centers had less than or equal to 4 subjects, which might be considered beyond random and suggest some qualitative difference. If the interaction is due to the poor planning or execution of the clinical trial in those centers, it is expected to be reduced, at least to the level of other cases, by standardization and training. There is little difference in the analysis of the Brief Psychiatric Rating Scale (BPRS) between the two groups.

The ICH E9 guideline suggests a procedure

TABLE 4

Subgroup Analysis for Center #16

	a	b	n	m	$\widehat{\sigma^2_\beta}/\widehat{\sigma^2_e}$	$m\widehat{\sigma^2_{\alpha\beta}}+\widehat{\sigma^2_e}/\sigma^2_e$	p	UC1	UC2
1. FGIR									
Large centers	2	5	146	14.1	0.036	0.835	0.51	3.12	1.15
Small centers	2	12	116	2.96	0	2.93	0.00	5.88	2.65
2. BPRS									
Large centers	2	5	146	14.1	0.007	0.931	0.45	3.54	1.18
Small centers	2	12	116	2.96	0	0.885	0.56	1.78	1.26

that tests the treatment effects first, ignoring interaction effects. However, assuming a fixed effects model as in this guideline and in the presence of interaction, the main effects may or may not appear according to their definition and lose their definite meaning; see Scheffé (8) and also the controversies in Senn (7). Having the result depend on the number of subjects recruited by each center seems to provide very little information since this would vary in the future population to be treated according to the distribution of those numbers. The usual procedure recommended in the textbooks is, therefore, testing the interaction first and if it is considered to exist, to then stop testing the main effects and begin interpreting the interaction. The guideline also recommends testing the interaction after detecting the main effects and then going back to interpret the main effects according to the interaction detected. Results of both approaches should, therefore, be essentially the same but the usual approach seems logically more acceptable.

It should further be noted that in the fixed effects model a contradiction exists between the sample size needed to demonstrate the absence of interaction using an equivalence test (9) and the sample size for demonstrating efficacy. A priori a compromise must be found that takes the different thresholds of center similarity and relevance of efficacy into account.

PLACEBO CONTROL

The ICH E9 guideline states that scientifically, efficacy is most convincingly established by: demonstrating superiority to placebo in a placebo controlled trial, showing superiority to an active control treatment, or demonstrating a dose-response relationship. It is very difficult to pass the superiority test against an active control; this is planned only when the test drug is definitely considered better than the active control. Therefore, superiority trials will primarily be conducted against placebo.

The placebo control is certainly justified in some cases, such as when developing a drug for nonresponders to established drugs. It is sometimes pointed out that without placebo researchers may not always be serious about every aspect of planning, practice, evaluation, and analysis of a trial since the incentive to show a clear difference between the test and control drugs is decreased. On the other hand, the use of placebo will be unethical when there is an established drug whose efficacy has been proven. In particular, in Japan where a clinical trial is conducted in the stream of usual clinical treatment, everyone has some kind of the public health insurance, and thus people have nothing to gain by participating in placebo controlled trials. Doctors have noted severe difficulties in obtaining informed consent from potential trial subjects. Paying an honorarium for participation in a trial has been forbidden in Japan.

Further, the following points regarding placebo controls should be considered seriously:

1. In Japan, people who respond to placebo are easier to recruit, resulting in an upper bias toward the placebo effect,
2. Seriously ill subjects are easier to exclude, causing bias if therapeutic improvement is dependent on the initial condition of the illness, in other words, if a larger improvement is expected for a more serious subject (10,11). In some cases, it is easier to recruit people who respond to the active agent, and
3. Blinding is easy to break.

Doctors in Japan have now begun to try to conduct more placebo controlled trials when necessary, and to recruit patients through newspaper advertisements. However, the above mentioned problems exist, and even if they were overcome, conducting only placebo controlled trials will not suffice for marketing if an established treatment exists; the ICH E9 guideline does not state this.

We do not want a drug whose relative usefulness against an existing drug has not been proven. This might, however, be specific to Japan where the price of a new drug is determined by the government relative to the price of a marketed drug in consideration of the relative potencies of the two drugs, and the consumer screening process of marketed drugs does not work well because everyone has public health insurance. This should be very different from the

United States, for example, where the price is under the severe surveillance of insurance companies.

Then naturally the concept of the three-arm trial with both placebo and active control arises. In this case, the demonstration of superiority of the test and standard treatments against placebo is required in addition to proving the noninferiority of the test treatment to the standard treatment. Although this is the simplest case among the more complicated cases considered, for example, by D'Agostino and Heeren (12), Dunnett and Tamhane (13), and Bauer et al. (14), this case still naturally loses power when there is a limited number of subjects to detect differences among treatments, compared to a trial with a simple hypothesis. The complicated cases discussed by D'Agostino and Heeren (12), Dunnett and Tamhane (13), and Bauer et al. (14) basically assume a normal model and require a relatively smaller sample size.

The problem of power is more concerned with efficacy rates. Due to lack of a sufficient number of patients in Japan, three-arm trials have rarely been conducted. This problem is partly due to the placebo control problem, as well as simply having fewer patients in Japan, as compared, for example, to the United States. Researchers in Japan are very eager to conduct bridging trials or international trials with a common protocol. Another point is that the endpoint for the active and placebo controls might be different; that is, a specific endpoint might be employed for the active control whereas a more general endpoint would be employed for the placebo. Careful conduct of three-arm trials is recommended.

Finally, in conducting dose-response trials, the range of doses is a major concern. It is easier to demonstrate a dose-response relationship by adopting a wide range of doses, probably beyond the optimal dose, and assigning more subjects to extreme dose levels (15) but this does not necessarily provide sufficient information on the dose to be used in usual treatment.

In Japan, phase 2 trials have been conducted with the same clinical endpoint as phase 3 trials, with the intention of finding the optimal effective dose for usual treatment. This makes it rather difficult to obtain definite evidence of the dose-response relationship. A significant dose response generally depends on variance, dose spacing, sample size allocation, number of doses, and shape of the dose-response; it is not a simple one-dimensional answer to efficacy. Including a group with a very low dose increases the chance of a significant dose-response but this might be unethical, like including a placebo group.

This might primarily be problematic due to the lack of sufficient patients and there might be no major conflict between these two attitudes of demonstrating a significant dose-response and finding an optimal dose in cases where a sufficient number of patients is available. It should, however, also be noted that using the dose levels whose efficacy and safety have been proven in clinical trials is strictly regulated in Japan and it is very difficult to search for an optimal dose after marketing. The simultaneous estimation of a dose which is both effective and safe using the intersection-union testing principle was recently proposed for the analysis of randomized dose finding studies (16).

NONINFERIORITY TEST AGAINST AN ACTIVE CONTROL

On the noninferiority test the two problems proposed by Temple (17) are always quoted:

1. How can the efficacy of the active control be ensured? and
2. How can a decrease in incentives for good clinical practice be prevented?

Much discussion has been generated regarding ensuring the efficacy of the active control. This has been convincingly settled by comparing the active control to placebo during drug development. Regarding decreased incentives for good clinical practice, the rule of nonsignificance has been replaced by a new procedure which requires in the case of the binomial distribution model, for example, the lower confidence bound of the difference $p_t - p_c$ not to go down below $-\Delta$ for some prespecified positive value Δ, where p_t and p_c denote the efficacy rates of the test and control drugs, respectively (18).

The ICH E9 guideline states that the new procedure is still not conservative. This is true in considering that the negative difference might be estimated to be positively biased, as in the case of intent-to-treat analysis. However, the following example shows that the procedure is much more reserved than it is perceived to be due to $-\Delta$ even if Δ is as much as 0.1 and the one-sided significance level is taken at $\alpha = 0.05$. Careful conduct of clinical trials is required.

Although there is naturally some attempt to change Δ according to the expected efficacy rate (19), the fixed Δ of 0.1 used to be employed in the European Union and Japan except for in cases of an extremely high efficacy rate beyond 0.9. We, therefore, use $\Delta = 0.1$ in the following discussions.

Suppose $p_t = p_c = 0.6$ and $\Delta = 0.1$. To ensure a power of 0.8 for proving noninferiority using an asymptotic test, the required sample size for one arm is 297 and 377 for $\alpha = 0.05$ and $\alpha = 0.025$, respectively. Those numbers are 198 and 252 when $p_t = p_c = 0.8$; this seems a little too large, at least compared to typical Japanese trials as shown earlier and in Table 3. If $p_t = 0.75$ and $p_c = 0.80$, then the required sample size is 862 and 1095 for $\alpha = 0.05$ and $\alpha = 0.025$, respectively, and such a trial will not be conducted unless there is a particular reason for doing so. Further, by the zero outcome of the difference ($\hat{p_t} - \hat{p_c} = 0$ by an obvious notation) in a trial with 100 subjects in one arm and $\Delta = 0.1$, only $p_t = p_c$ above 0.75 and 0.85 will be declared to be noninferior at $\alpha = 0.05$ and, $\alpha = 0.025$, respectively.

These considerations show that negative trials where $p_t < p_c$ have very little chance of being conducted and clearing the lower bound, despite Temple's comment (17), even if $\alpha = 0.05$ and $\Delta = 0.1$. It should be noted that $\Delta = 0.1$ does not imply that the efficacy rates of the outgoing drugs passing through these tests is $p_c - 0.1$. They will be much higher than $p_c - 0.1$ in trials of the usual scale.

Japan previously used the one-sided significance level of 0.05. The ICH E9 guideline unifies the one-sided significance level at 0.025. It is difficult to discuss which is appropriate since there is no theoretical basis to define α. Instead, α should be defined based on experience. Further discussions will be necessary on the resulting outgoing efficacy rate for drugs that pass through noninferiority tests.

Japan has had experience with a one-sided significance level of $\alpha = 0.05$ related to the noninferiority test with a handicap of $\Delta = 0.10$ since the previous guideline was issued in 1992 (18). It is also considered inappropriate to use α as a tuning variable when there is another tuning variable Δ. Instead, there is an interesting statistical logic to naturally combine one- and two-sided tests if we design all the tests to be of the same significance level α (20).

Let the parameter space of $p_t - p_c$ be partitioned into three parts,

$$H_1 : p_t - p_c > 0,$$

$$H_2 : p_t - p_c = 0,$$

$$H_3 : p_t - p_c < 0.$$

Then according to the test with significance level α for each of the hypotheses H_1, H_3 (one-sided), and H_2 (two-sided), a confidence region for $p_t - p_c$ with a confidence coefficient of $1 - \alpha$ is formed as follows:

$$\begin{aligned}
K_{\alpha/2} &< (\hat{p_t} - \hat{p_c}) / G^{1/2} && \rightarrow p_t > p_c \\
K_{\alpha} &< (\hat{p_t} - \hat{p_c}) / G^{1/2} \leq K_{\alpha/2} && \rightarrow p_t \geq p_c \\
-K_{\alpha} &\leq (\hat{p_t} - \hat{p_c}) / G^{1/2} \leq K_{\alpha} && \rightarrow p_t \gtreqless p_c \text{ (no choice)} \\
-K_{\alpha/2} &\leq (\hat{p_t} - \hat{p_c}) / G^{1/2} < -K_{\alpha} && \rightarrow p_t \leq p_c \\
& (\hat{p_t} - \hat{p_c}) / G^{1/2} < -K_{\alpha/2} && \rightarrow p_t < p_c .
\end{aligned}$$

Here we assumed a normal approximation although a more elaborate discussion is possible, and K is the upper α point of the standard normal distribution and $G = (n_c^{-1} + n_t^{-1})\,\hat{p}(1 - \hat{p})$ is the variance estimator for the estimate $\hat{p}_t - \hat{p}_c$ with n_t and n_c the sample sizes for the test and control drugs, respectively; p is the maximum likelihood estimator of $p_t = p_c$ under H_2. This procedure can also be interpreted as the special case of the closed testing procedure (21) where the intersection of any two hypotheses among H_1, H_2, and H_3 is empty. Then we can add $H_4 : p_t - p_c <-\Delta$ and test it at the same significance level α. Since it is included in H_3, we have $H_4 \cap H_3 = H_4$, $H_4 \cap H_1 = H_4 \cap H_2 = \varnothing$ (empty set). We, there-

fore, test H_4 first and according to the closed testing procedure we proceed to test other hypotheses only when H_4 is rejected. Then we can summarize the result as follows.

1. If H_4 is not rejected we conclude that noninferiority of the test drug against the control drug cannot be confirmed,
2. If H_4 is rejected but H_3 is not, then we can only assert the noninferiority: $p_t \geq p_c - \Delta$,
3. If H_3 is rejected but H_2 is not then we can assert at least equivalence: $p_t \geq p_c$, and
4. If H_2 is rejected in favor of H_1, namely $G^{-1/2}(\widehat{p_t} - \widehat{p_c}) > K_{\alpha/2}$, we can assert the superiority of the test drug against the control: $p_t > p_c$.

All the tests in the procedure, either one-sided or two-sided, are performed at the same significance level α. This procedure thus combines the noninferiority and superiority tests and one- and two-sided tests very naturally by changing the strength of the evidence obtained of the goodness of the test drug against the control. It should be noted that in this procedure we need not prespecify the distinction of noninferiority and superiority tests or one- and two-sided tests, thus answering the frequent questions about how to argue the significance of superiority results obtained under noninferiority trials. See also Morikawa and Yoshida (22) for the combined tests of superiority and test of equivalence.

Ethics is another issue to be discussed regarding noninferiority tests. In the old Japanese guideline, noninferiority was argued as the necessary condition for a drug which has another advantage over the standard drug, such as safety, ease of administration, long shelf life, and so on. However, in the ICH E9 guideline, the noninferiority test is argued to be only parallel to the superiority test without any reference to additional advantages. Thus, the concept of noninferiority naturally invites an ethical discussion just as the use of placebo control does, especially in Japan where trials are conducted in the stream of usual clinical treatment and everyone has some kind of public health insurance.

In Japan, therefore, the priority of the superiority test over the noninferiority test is eagerly argued (23). Then, however, the difficulty in proving superiority should be taken into consideration. Suppose that the efficacy rates of the test and control drugs are 0.80 and 0.70, respectively. Then the necessary sample size for one arm to ensure a power of 0.8 in the superiority test with a two-sided $\alpha = 0.05$ is as large as 294 by asymptotic theory for a simple binomial model. This is again beyond the usual scale of Japanese trials and usually invites a discussion about the availability of patients. If the difference amounts to twice that of the above example, as can be expected in the case of placebo control, the necessary sample size goes down approximately to one fourth and becomes feasible.

On the other hand, by the noninferiority test with $\Delta = 0.1$, the necessary sample size is 58 and 74 for each of the one-sided $\alpha = 0.05$ and $\alpha = 0.025$, respectively. If the efficacy rates are 0.75 and 0.70 for the test and control drugs, respectively, the necessary sample size goes up to 1251 for the superiority test with two-sided $\alpha = 0.05$, compared to 110 and 139 for the noninferiority test with $\Delta = 0.1$ and one-sided $\alpha = 0.05$ and $\alpha = 0.025$, respectively. These considerations suggest that the superiority test with two-sided $\alpha = 0.05$ is often too reserved and will lack power at the usual scale of phase 3 trials in Japan. Thus, some sort of noninferiority test would be necessary. The efficacy rates of the outgoing drugs would be much higher than the impression received from the word noninferiority and the maximal tolerance $-\Delta$. We would, therefore, like to invite a lot of simulation work, in the realistic situation of the respective application field, on the affect of Δ and α on the outgoing efficacy rate.

SUBJECTIVE AND OBJECTIVE MEASURES

The clinical evaluation of a subject is essentially multivariate and only in the best case do we have a single quantitative primary endpoint. As shown in the earlier examples, doctors in Japan have historically used a global assessment variable based on those multiple measurements when there is a single quantitative primary endpoint. In recent years, however, discussions have

been more in favor of objective measures, such as physical measurements or universally accepted rating scales such as Hamilton Anxiety Scale (HAM-A), Hamilton Depression Scale (HAM-D), and BPRS. As a reaction to the past, some Japanese doctors even seem to have faith in quantitative measures. It is, therefore, worthwhile to compare the advantages and disadvantages of objective and subjective measures.

PROBLEMS WITH OBJECTIVE MEASURES

Linearity, Additivity, and Normality. Even with a quantitative measurement, it is not clear if linearity, additivity, or normality are satisfied. Clinical measurements are often nonlinear. Without an appropriate transformation, they are quite misleading, however, appropriate transformations are not obvious. To lower blood pressure by 30 mm Hg from an initial value of 220 mm Hg is clinically different, for example, than lowering blood pressure by 30 mm Hg from 170 mm Hg. We must often plot the change in the measurement against the initial value to determine whether to use the change itself or the ratio of the change to the initial value. If this problem is not apparent, it is simply because the range of the initial value is restricted in recruiting patients in the trial. There is also discussion on the J-shape effect with blood pressure, as explained later (24,25).

Further, some statistical methods are sensitive to outliers or nonnormality, which is very common in clinical measurement. We see many examples where the distribution of variables is lognormal and the normal theory fails to prove the difference between the treatments by the overestimated error variances or erroneously detects outliers that are not abnormal under the lognormal distribution. In that case, quantitative measure is not necessarily more informative than the rank or the ordered categorical data.

Measurement Errors. Objective measurements also suffer from various sources of measurement errors. First, there are errors related to a subject, such as the circadian rhythm of blood pressure, hypertension in the presence of a doctor, and variations in total cholesterol due to the food consumed before measurement. Even for the exact measurements such as blood pressure or bone mass, variations due to the experimenter, instrument, and institute are unexpectedly large beyond the range of random variation. To make those measurements more reliable, standardization and training is strongly recommended.

Difficulty Summarizing Quantitative Measurements. Quantitative measurements are usually obtained as multivariate variables. Even for blood pressure, which is regarded as a typical quantitative measure, the appropriate primary endpoint is unclear; it could be: systolic blood pressure, diastolic blood pressure, or the average measurements of those two; morning, daytime, or night measurements; the difference between the initial and final measurements; the ratio of the difference to the initial value; trough peak ratio; binary response as to whether blood pressure is normally controlled, and so on. At one time, the trough peak ratio was recommended as the primary endpoint but problems related to this were recognized. More recently, the lowering effect measured simply by the difference between the initial and resulting values after treatment has been gaining favor. This should be acceptable in the trial when the range of patients is restricted, at the cost of reducing generalizability.

A large-scale clinical trial is underway in Japan to verify whether the same criterion can be applied to senior citizens. Similarly, there are several proposals with respect to cholesterol measurements such as LDL, non-HDL, TC-HDL, or TC/HDL. Further, there is an objection to using a simple total sum of scores of the rating scales with different responsiveness to the agent in the anti-depressant drug (11). It is, therefore, urgent to achieve consensus on the primary endpoint in each disease.

Difficulty Comparing Repeated Measurements. Blood pressure, cholesterol, and bone mass measurements are obtained as repeated measures for several months or years. The analysis, however, is often based on the initial and fi-

nal values only and a more elaborate profile analysis will be required. This is not a criticism of the quantitative measurements; it refers to the ad hoc procedures seen in the usual practice.

Agreement with the Clinical Endpoint. Quantitative measures are useful not because of their objectivity but because of their agreement with the clinical endpoint. It is, therefore, necessary to approve every quantitative surrogate endpoint to be truly useful by a large-scale clinical trial. Regarding blood pressure, for example, the J-shape effect of diastolic blood pressure on cardiovascular disease has been recognized (24), that is, there is an assertion of optimal value around 85 ~ 90 mm Hg. Discussion on this is ongoing (25).

ADVANTAGES AND DISADVANTAGES OF GLOBAL ASSESSMENT

The ICH E9 guideline acknowledges or even recommends the use of global assessment variables in some cases to measure the overall safety, efficacy, and/or usefulness of a treatment. Global assessment variables integrate objective measurements and the investigator's overall impression about the subject's state or change in state and thus, inevitably have a subjective component. In Japan, however, as a reaction to too frequent use of global assessment variables in the past, there is a tendency to refrain from using them by instead using quantitative measurements. It is time to discuss when, where, and how to appropriately use global assessment variables.

As stated in the ICH E9 guideline, the relevance of the subjective scale to the primary objective of the trial, and the process used to integrate the collected quantitative measures and the investigator's impression, should be mentioned in the protocol. However, it should not be too strict (eg, like a mathematical equation) since there is a subjective component to global assessment.

The disadvantage of global assessment is certainly its large variations among doctors due to its subjective nature (a sort of measurement error). The advantages of global assessment are:

1. It is a summary measure that does not need a sophisticated multivariate analysis and is regarded as a surrogate marker of quality of life since it takes the overall condition of subjects into account,
2. It can adapt to nonlinear clinical responses, to some extent,
3. It considers the time profile, initial condition of illness, and various personal aspects of the subjects, and
4. It can be used to some extent when there are missing values, which are inevitable in clinical trials.

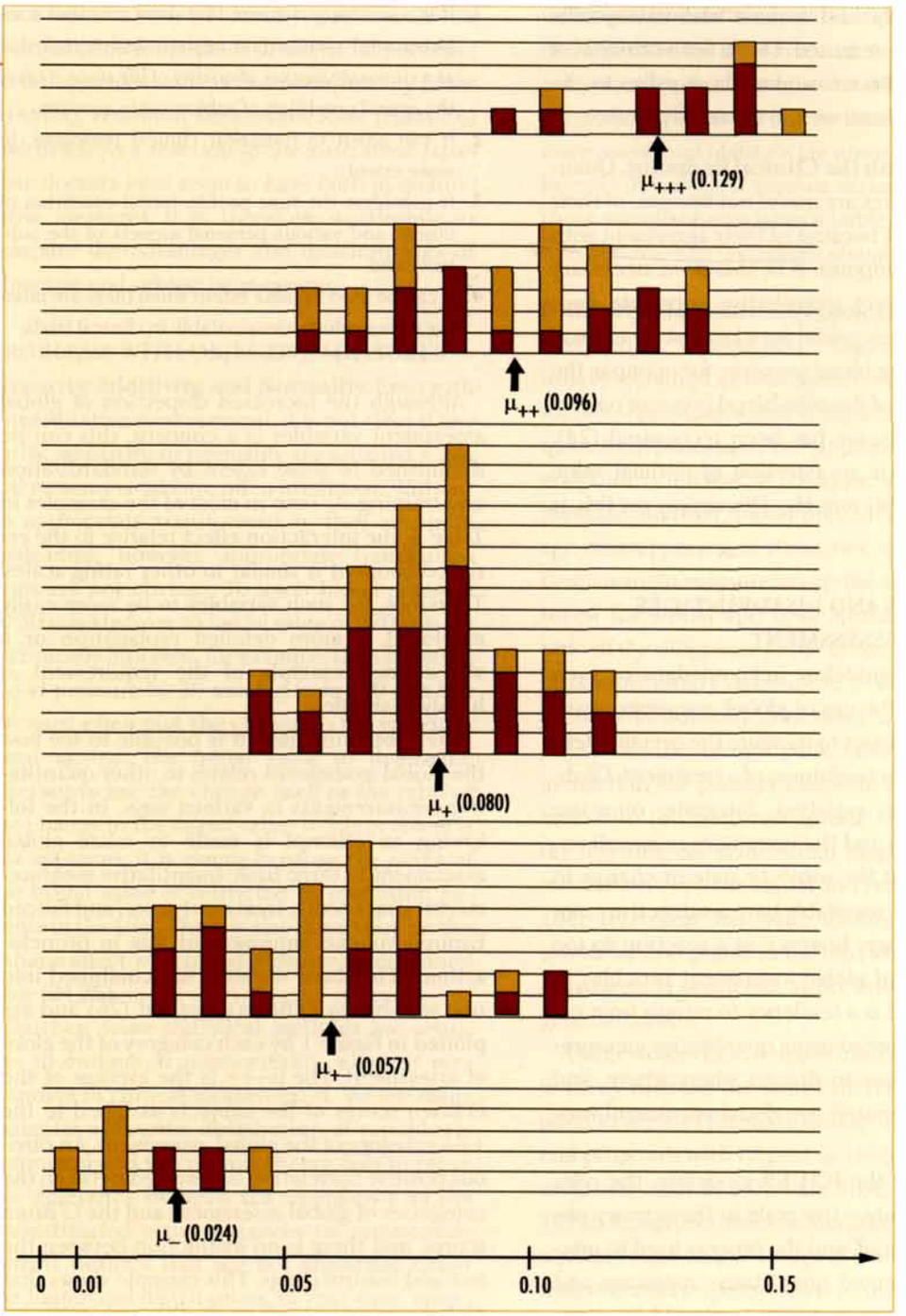
Although the increased dispersion of global assessment variables is a concern, this can be diminished to some extent by standardization and training. As seen in most of the examples in Table 3, the interaction effect relative to the error component is similar to other rating scales. Therefore, for such variables to be more easily employed, a more detailed explanation or a well-defined example of the requirement is highly desirable.

After obtaining data, it is possible to see how the global assessment relates to other quantitative measurements in various ways. In the following an attempt is made to relate global assessment to three basic quantitative measurements: attack score, treatment score, and Forced Expired Volume improvement rate in bronchiasthma. The three variables are combined into one variable by O'Brien's method (26) and are plotted in Figure 1 by each category of the global assessment. The μ+++ is the average of the O'Brien scores of the subjects assigned to the +++ category of the global assessment. An obvious positive correlation is observed between the categories of global assessment and the O'Brien scores, and there is no distinction between the test and control drugs. This example shows that global assessment will work, although some increase of sample size is necessary due to the large variations in global assessment. Standardization and training in the process of forming global assessments will reduce those variations.

In the second example for the same disease, shown in Figure 2, something strange occurs. If the O'Brien scores for the placebo and the test drug are the same, there is a tendency for the

FIGURE 1

Distribution of O'Brien's score at each category of the global assessment (red: test drug; gold: active control).



placebo to be assigned to lower categories. This resulted in the absence of placebo in the highest category and the distribution of placebo in ++ and + categories shifted upward, compared to those of the test drug. This is an old example of a trial and the only quantitative measurements are the attack and treatment scores; thus, the global assessment and the quantitative measurements might be measuring different things. This is, however, only one attempt. Every effort to relate the global assessment to quantitative measures should be made.

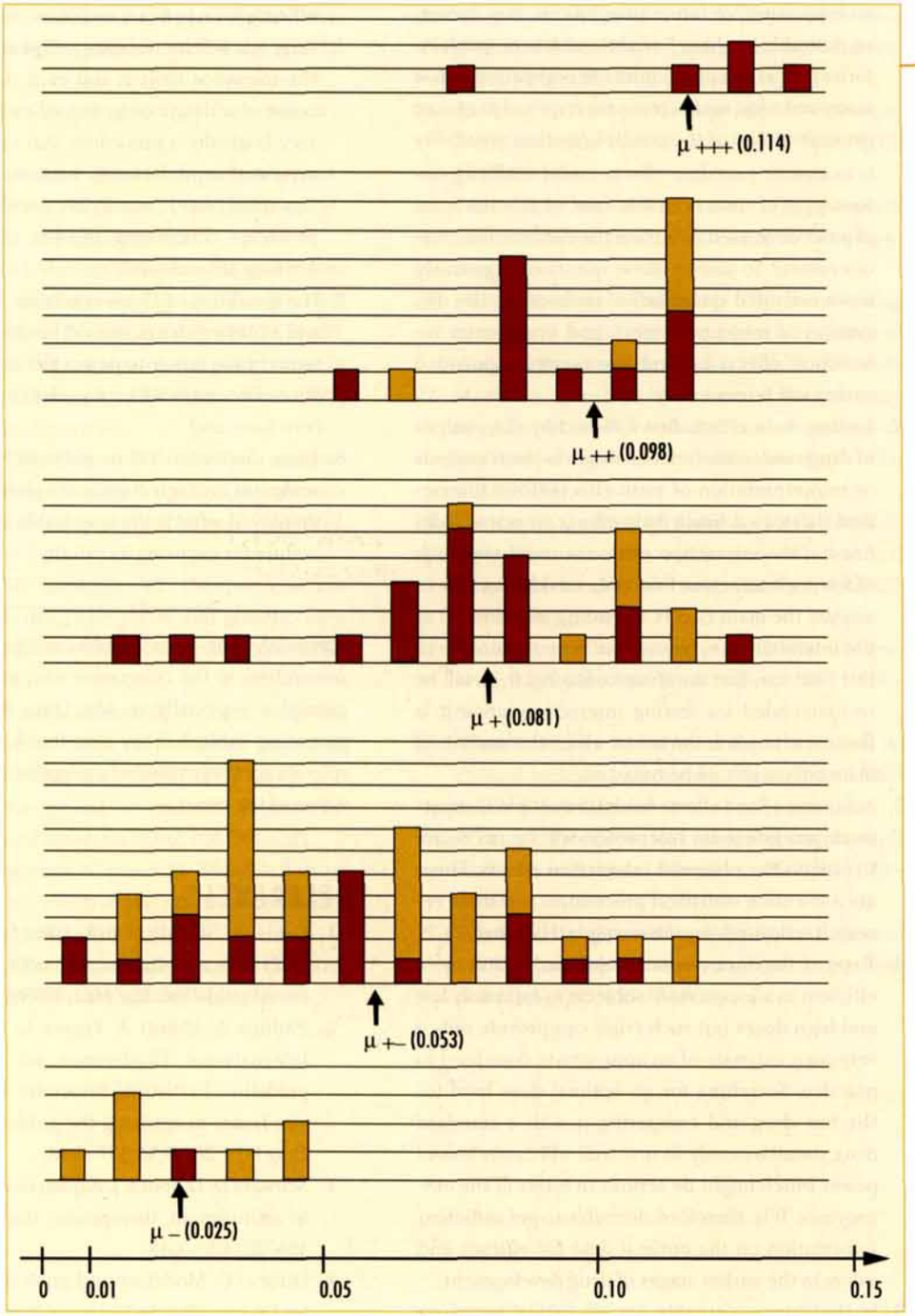


FIGURE 2

Distribution of O'Brien's score at each category of the global assessment (red: test drug; gold: placebo).

CONCLUSION

The ICH E9 guideline's emphasis on the need for source control and robust design to minimize the various sources of bias and to obtain robust conclusions is welcome. As the guideline suggests, the trial endpoints, target population, design and sample size, statistical hypothesis, and proposed statistical method, as well as its reasoning based on past research, must be specified clearly in the protocol. In this regard, the following points should be discussed more intensively:

1. How many and what type of centers should be involved in a trial? By using a fixed effects model for centers hypothesis testing is intended rather than

an estimation for future populations. It is, however, desirable in phase 3 trials to obtain as much information as possible for future populations. How many and what sort of trials are required to ensure generalizability of the results? Another possibility is to assume a random effects model involving various types of centers. In this case, what is the actual procedure used to ensure the random selection of centers? To answer these questions rigorously, more extended quantitative analyses on the dispersion of subjects, centers, and drug-center interaction effects beyond the examples provided earlier will be necessary,

2. Testing main effects first followed by the analysis of drug-center interactions requires the reanalysis or reinterpretation of main effects if the interaction is detected. Since main effects are not well defined, if the interaction exists the usual approach of testing interaction first and considering how to analyze the main effects according to the result of the interaction analysis seems more reasonable in this case too. The significance level of 0.15 will be recommended for testing interaction since it is the test to confirm the model which the analysis of main effects should be based on,
3. Assuming a fixed effects model, a multiple comparisons procedure for interaction will be necessary to explain the observed interaction effects. There are a few such statistical procedures and more research is desired, see, for example, Hirotsu (27),
4. To prove the dose-response relationship, it is more efficient to allocate more subjects to extremely low and high doses but such trials can provide only a very poor estimate of an appropriate dose level in practice. Searching for an optimal dose level for the test drug and comparing it with a standard drug simultaneously in one trial will cause loss of power which might be serious in terms of the efficacy rate. It is, therefore, desirable to get sufficient information on the optimal dose for efficacy and safety in the earlier stages of drug development,
5. Is the drug acceptable for marketing based on passing the superiority test against placebo but without any comparison to standard marketed drugs? It seems desirable to have some idea of the relative efficacy of the test drug against the standard drug before marketing since it is difficult to compare drugs on the market properly due to various noise factors,
6. A superiority test against an active control is too difficult unless there is a prominent difference in efficacy for a primary endpoint; thus, a noninferiority test will be necessary. Research determining the tolerance limit Δ and even α from the viewpoint of a drug's outgoing efficacy rate is necessary. Logically, a procedure that combines superiority and noninferiority tests and also one- and two-sided tests is worthy of consideration. Ethical problems concerning the use of noninferiority tests are also relevant,
7. The feasibility of three-arm trials in test, placebo, and standard drugs should be discussed more in terms of the power to detect the intolerable difference. The availability of patients is a major concern here, and
8. More discussion will be necessary on the primary endpoint for each disease. If a global assessment is employed, what is the reasonable and feasible procedure for ensuring its validity?

Acknowledgment—The authors are greatly obliged to researchers in the companies who provided the real examples, especially to Ms. Ohta for her help in preparing Table 3. They also thank the anonymous referees for their valuable comments which led to revision of the paper.

REFERENCES

1. Lewis JA. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. *Stat Med.* 1999;18:1903–1942.
2. Phillips A, Ebbutt A, France L, Morgan D. The International Conference on Harmonization guideline 'Statistical Principles for Clinical Trials': Issues in applying the guideline in practice. *Drug Inf J.* 2000;34:337–348.
3. Schwarz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chron Dis.* 1967;20:637–648.
4. Hirotsu C. Modeling and analyzing the generalized interaction. In *Statistical Monitoring and Optimization.* Parla SH, Vining GG, eds. New York, NY: Marcel Dekker; 2000.
5. Gould AL. Multi-center trial analysis revisited. *Stat Med.* 1998;17:1779–1797.
6. Hirotsu C. *Estimating variance components in a two-way layout with unequal numbers of observation. Report of Statistical Application and Research.* Tokyo, Japan: Japanese Union of Science and Engineering. 1966;13(2):29–43.

7. Senn S. Some controversies in planning and analyzing multi-center trials. *Stat Med.* 1998;17: 1753–1765.
8. Scheffé H. *The Analysis of Variance.* New York, NY: Wiley; 1959.
9. Wellek S. Testing for absence of qualitative interactions between risk factors and treatment effects. *Biomet J.* 1997;39:809–821.
10. Uhlenhuth EH, Matuzas W, Warner TD, et al. Growing placebo response rate: The problem in recent therapeutic trials? *Psychopharm Bull.* 1997;33:31–39.
11. Ishigooka J. A critical view on the placebo use in clinical trials. *Japanese J Clin Psychopharm.* 1999; 2:145–153 (in Japanese).
12. D'Agostino RB, Heeren TC. Multiple comparisons in over the counter drug clinical trials with both positive and placebo controls. *Stat Med.* 1991;10:1–6.
13. Dunnett CW, Tamhane AC. Comparisons between a new drug and active and placebo controls in an efficacy clinical trial. *Stat Med.* 1992; 11:1057–1063.
14. Bauer P, Röhmel J, Maurer W, Hothorn L. Testing strategies in multidose experiments including active control. *Stat Med.* 1998;17:2133–2146.
15. Hirotsu C. An optimal design for an isotonic inference. *J Stat Plann Infer.* 2002;106:205–213.
16. Bauer P, Brannath W, Posch M. Multiple testing for identifying effective and safe treatments. *Biomet J.* 2001;43:605–616.
17. Temple R. Difficulties in evaluating positive control trials. *Clin Eval.* 1993;21:141–149.
18. Kouseishou. *A Guideline for the Statistical Analysis of Clinical Trials.* Tokyo, Japan: Japanese Ministry of Public Health; 1992 (in Japanese).
19. Röehmel J. Therapeutic equivalence investigations: statistical considerations. *Stat Med.* 1998; 17:1703–1714.
20. Takeuchi K. *Methodological Basis of Mathematical Statistics.* Tokyo, Japan: Toyokeizai;1973 (in Japanese).
21. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika.* 1976;63:655–660.
22. Morikawa T, Yoshida M. A useful testing strategy in phase III trials: combined test of superiority and test of equivalence. *J Biopharm Stat.* 1995; 5:297–306.
23. Hirotsu C, Kurihara M, Shimizu N, et al. Roundtable discussion on application of statistical principles for clinical trials. *Clin Eval.* 1999;27: 13–66 (in Japanese).
24. Cruickshank J. Benefit and potential harm of lowering high blood pressure. *Lancet.* 1987;1:581–584.
25. Goto E, Moriya A. J shaped hypothesis and its clinical meaning. *Circulatory Orgom Today.* 1997; 1:312–316 (in Japanese).
26. O'Brien P. Procedure for comparing samples with multiple endpoints. *Biometrics.* 1984;40:1079–1087.
27. Hirotsu C. An approach to defining the pattern of interaction effects in a two-way layout. *Ann Instit Stat Math.* 1983;A35:77–90.