PLOS ONE

RESEARCH ARTICLE

Sample size determination for a specific region in multiregional clinical trials with multiple co-primary endpoints

Wong-Shian Huang[1,2], Hui-Nien Hung[1], Toshimitsu Hamasaki[3], Chin-Fu Hsiao[1,2]*

1 Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan, **2** Institute of Population Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan, **3** National Cerebral and Cardiovascular Center, Osaka, Japan

* chinfu@nhri.org.tw

OPEN ACCESS

Citation: Huang W-S, Hung H-N, Hamasaki T, Hsiao C-F (2017) Sample size determination for a specific region in multiregional clinical trials with multiple co-primary endpoints. PLoS ONE 12(6): e0180405. https://doi.org/10.1371/journal.pone.0180405

Editor: Tim Friede, University Medical Center Gottingen, GERMANY

Received: July 24, 2016

Accepted: June 15, 2017

Published: June 30, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The work in this paper was supported by the grant MOST 103-2118-M-400-004- from the Ministry of Science and Technology, Taiwan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Recently, multi-regional clinical trials (MRCTs), which incorporate subjects from many countries/regions around the world under the same protocol, have been widely conducted by many global pharmaceutical companies. The objective of such trials is to accelerate the development process for a drug and shorten the drug's approval time in key markets. Several statistical methods have been purposed for the design and evaluation of MRCTs, as well as for assessing the consistency of treatment effects across all regions with one primary endpoint. However, in some therapeutic areas (e.g., Alzheimer's disease), the clinical efficacy of a new treatment may be characterized by a set of possibly correlated endpoints, known as multiple co-primary endpoints. In this paper, we focus on a specific region and establish three statistical criteria for evaluating consistency between the specific region and overall results in MRCTs with multiple co-primary endpoints. More specifically, two of those criteria are used to assess whether the treatment effect in the region of interest is as large as that of the other regions or of the regions overall, while the other criterion is used to assess the consistency of the treatment effect of the specific region achieving a pre-specified threshold. The sample size required for the region of interest can also be evaluated based on these three criteria.

Introduction

Recently, global drug development has attracted much attention from pharmaceutical companies. Unlike traditional clinical trials, the design of MRCT recruiting subjects from many countries around the world under the same protocol has led to a new strategy for drug development. This kind of design has been widely adopted by global pharmaceutical companies, which seek simultaneous drug development, submission, and regulatory approval throughout key world markets to hasten the market availability of the drug, as well as improved patient access to new and innovative treatments. However, a key issue for conducting MRCTs is how to demonstrate the efficacy of a drug in all participating regions while also evaluating the

possibility of applying the overall trial results to each region. To address the difficulties related to global drug development, in 1998 the International Conference on Harmonization (ICH) published "Ethnic Factors in the Acceptability of Foreign Clinical Data", known as the E5 guideline. The idea of an MRCT was first raised in the 11th Q& A of E5 [1]. In recent years, the trend for simultaneous clinical development in the world has been rapidly rising. To establish a framework for how to demonstrate the efficacy of a drug in all participating regions while also evaluating the possibility of applying the overall trial results to each region by conducting an MRCT, the ICH released the draft E17 guideline "General principle on planning/designing Multi-Regional Clinical Trials" [2] in 2016 to describe general principles for the planning and the design of MRCTs; another aim of the work was to increase the acceptability of MRCTs in global regulatory submissions.

The Japanese Ministry of Health, Labour and Welfare issued its own guidance document on MRCTs, "Basic Principles on Global Clinical Trials" [3]. This guidance provided two methods as examples to determine the number of Japanese subjects required for establishing consistency in treatment effect between the Japanese group and the entire group. Let D_{Japan} and D_{All} represent the observed treatment effects for the Japanese group and the entire group. Method 1 in the Japanese guidance suggests that the sample size for Japan should fulfill

$$P(D_{\text{Japan}}/D_{\text{All}} > \pi) > \gamma.$$

On the other hand, suppose that an MRCT will be conducted in three regions. Let D_i represent the observed treatment effect for region i, $i = 1,\ldots,3$. For Method 2, the sample size should be determined to satisfy

$$P(D_1 > 0, D_2 > 0, D_3 > 0) > \gamma.$$

Note that the Japanese guidance requires that π is 0.5 or greater and that γ be 0.8 or greater.

Several different statistical approaches based on Methods 1 and 2 in the Japanese guidance have been developed. Quan et al. [4] calculated the sample size required for Japan in an MRCT with normal, binary, and survival endpoints based on Method 1. Kawai et al. [5] proposed an approach, based on Method 2, to allocate the total sample size to the regions so that a high probability of observing a consistent trend under the assumed treatment effect across regions can be obtained. In addition, consistency criteria different from those of the Japan guidance have been established, such as those by Tsou et al. [6], Uesaka [7], Ko et al. [8], and Tsou et al. [9]. On the other hand, Chen et al. [10] and Huang et al. [11] considered ethnic differences and proposed methods that apply different treatment effects across regions to the design and evaluation of MRCTs.

However, most recent approaches to the design and evaluation of MRCTs are concerned with only one primary endpoint. In some therapeutic areas the clinical efficacy of a new treatment may be characterized by a set of possibly correlated endpoints, because there may be several different aspects to patients' responses to that treatment. For example, a typical clinical trial for Alzheimer's disease (AD) is usually conducted with cognitive, functional, and global endpoints to evaluate a symptomatic improvement in the dementia caused by the disease; the Committee for Medicinal Products for Human Use (CHMP) [12] and the Food and Drug Administration (FDA) [13] have recommended the two co-primary endpoints of these three in the development of drugs for the treatment of AD, where clinical trials with "co-primary" endpoints are designed to evaluate if the effect of a test treatment is superior (or non-inferior) to the control on all primary endpoints. Failure to demonstrate superiority on any single endpoint implies that superiority to the control treatment cannot be concluded. These endpoints are classified as follows:

1. objective cognitive tests, e.g., the AD Assessment Scale cognitive subscale(ADAS-cog) and Severe Impairment Battery (SIB);
2. self-care and activities of daily living, e.g., the AD Cooperative Study Activities of Daily Living (ADCS-ADL) and its modified version for severe AD; and
3. global assessment of change, such as the Clinician's Interview Based Impression of Change-plus (CIBIC-plus) and the Clinical Global Impression of Improvement (CGI-I).

Having such multiple endpoints raises difficulties for statisticians in handling multiplicity in the design and analysis of clinical trials, specifically controlling Type I and Type II error rates when the endpoints are potentially correlated. When designing a trial to evaluate joint effects on all endpoints, as seen in AD clinical trials, no adjustment is needed to control the Type I error rate. However, the Type II error rate increases as the number of endpoints to be evaluated increases. This situation is referred to as "multiple co-primary endpoints" and it is related to the intersection-union problem (Hung and Wang [14]; Offen et al. [15]).In many such trials, the sample size is often unnecessarily large, which results in complications. To overcome the issue, recently many authors have discussed approaching the design and analysis of co-primary endpoints trials using fixed-sample (size) design; the extensive references in Offen et al. [15] and Sozu et al. [16] provide many examples.

In this paper, we will focus on the design and evaluation of an MRCT with multiple co-primary endpoints. As we know, the aim of an MRCT is to show the efficacy of a drug in various global regions, and concurrently to evaluate the possibility of applying the overall trial results to each region. Therefore, we will also consider the determination of the number of subjects in a specific region to establish the consistency of treatment effects between the specific region and the entire group.

This paper is organized as follows. In section 2, we demonstrate the sample size calculation for multiple endpoints with correlation. In section 3, we established three criteria to assess the consistency of treatment effects between a specific region and the entire group in MRCTs with multiple endpoints. Under each criterion, the sample size required for the region of interest is also evaluated. An example is provided in section 4. Discussions are given in section 5.

Material and methods

Sample size calculation

For simplicity, we focus on a most fundamental situation, where an MRCT is designed to evaluate superiority over a placebo control on $K(\geq 2)$continuous multiple co-primary efficacy endpoints, and the effect size for each co-primary endpoint is assumed to be uniform across $M(\geq 2)$ regions. Consequently, we can let X_{ikj} and Y_{ikl} be efficacy responses on the kth co-primary endpoint for the jth subject and for the lth subject in the ith region receiving the test product and the placebo control, respectively, $i = 1,\ldots,M$, $j = 1,\ldots,N_i^{\mathrm{T}}$, $l = 1,\ldots,N_i^{\mathrm{C}}$, and $k = 1,\ldots,K$. Let $\mathbf{X}_{ij} = (X_{i1j}, X_{i2j},\ldots, X_{iKj})^T$ and $\mathbf{Y}_{il} = (Y_{i1l}, Y_{i2l},\ldots, Y_{iKl})^T$ be the outcome vectors of K co-primary endpoints for the jth subject and the lth subject in the ith region receiving the test product and the placebo control, respectively, $j = 1,\ldots,N_i^{\mathrm{T}}$, $l = 1,\ldots,N_i^{\mathrm{C}}$.

Since the effect size for each co-primary endpoint is uniform across regions, we can therefore assume that $\mathbf{X}_{ij}$ and $\mathbf{Y}_{il}$ have multivariate normal (MVN) distributions with population mean vectors $\mu^T = (\mu_1^T, \ldots, \mu_K^T)$ and $\mu^C = (\mu_1^C, \ldots, \mu_K^C)$, respectively, and a known common covariance matrix $\Sigma = (\rho_{kk'}\sigma_k\sigma_{k'})$, where $(a_{kk'})$ denotes the matrix whose $(k,k')^{\mathrm{th}}$ element is $a_{kk'}$, $\rho_{kk'} = corr(X_{ikj}, X_{ik'j}) = corr(Y_{ikl}, Y_{ik'l})$, $k \neq k'$, and $\sigma_k^2 = \mathrm{Var}(X_{ikj}) = \mathrm{Var}(Y_{ikl})$. Here, we assume that the outcome variances are known, although in actual practice, they are usually

unknown and must be estimated from some data. Let $\Delta_k = \mu_k^T - \mu_k^C$ for $k = 1,\ldots K$. Here a higher value of the population mean for each co-primary endpoint represents a better outcome. Consequently, the hypothesis testing for multiple co-primary endpoints is given as

$$\mathrm{H}_0 : \Delta_k \leq 0 \text{ for at least one k vs. } \mathrm{H}_\mathrm{A} : \Delta_k > 0 \text{ for all } k. \quad (1)$$

The null hypothesis H_0 can be conveniently expressed as a union of a family of hypotheses. The hypothesis for each co-primary endpoint is tested at the same significance level of α with H_{0k}:$\Delta_k \leq 0$ vs. H_{Ak}:$\Delta_k > 0$, and the null hypothesis $\mathbf{H}_0$ is rejected if and only if each null hypotheses $\mathbf{H}_{0k}$ is rejected, so that the hypothesis testing for multiple co-primary endpoints is a test of the significance level of α. Although the hypothesis is one-sided, the proposed method can be straightforwardly extended to the two-sided hypothesis. Let

$$\bar{X}_k = (\sum_{i=1}^{M}\sum_{j=1}^{N_i^T} X_{ikj})/(\sum_{i=1}^{M} N_i^T) \text{ and } \bar{Y}_k = (\sum_{i=1}^{M}\sum_{j=1}^{N_i^C} Y_{ikj})/(\sum_{i=1}^{M} N_i^C),$$

for $k = 1,\ldots K$. Also let

$$Z_k = \frac{\bar{X}_k - \bar{Y}_k}{\sigma_k \sqrt{\frac{1}{\sum_{i=1}^{M} N_i^T} + \frac{1}{\sum_{i=1}^{M} N_i^C}}},$$

for $k = 1,\ldots K$. Subsequently, we will reject H_0 at α level of significance if

$$Z_k > z_{1-\alpha} \text{ for all } k,$$

where $z_{1-\alpha}$ is the 100(1-α) percentile of the standardized normal distribution.

Let $N^T = \sum_{i=1}^{M} N_i^T$ and $N^C = \sum_{i=1}^{M} N_i^C$. In the design stage we assume equally sized groups, i.e., $N^T = N^C = N$. Let $\mathbf{Z} = (Z_1,\ldots,Z_K)^T$. Then, under H_1, $\mathbf{z}$ is distributed as an MVN with mean vector $(\sqrt{N/2})\boldsymbol{\delta}$ and covariance matrix $\boldsymbol{\rho} = (\rho_{kk'})$, where $\boldsymbol{\delta} = (\Delta_1/\sigma_1,\ldots,\Delta_K/\sigma_K)^T$.

Using the result in Sozu et al. [17,18], the power for rejecting the null hypothesis H_0 can be written as

$$1 - \beta = \Pr\left[\bigcap_{i=1}^{K}\{Z_k > z_{1-\alpha}|\mathrm{H}_A\}\right].$$

This power is referred to as "conjunctive power" (Senn and Bretz [19]) or "complete power" (Westfall et al. [20]). The sample size required for achieving the desired power of 1 $-\beta$ at the significance level of α for the one-sided test can be found by the minimum N that satisfies

$$\int_{z_{1-\alpha}}^{\infty} \cdots \int_{z_{1-\alpha}}^{\infty} f\left(z_1,\ldots,z_K;\left(\sqrt{N/2}\right)\boldsymbol{\delta},\boldsymbol{\rho}\right) dz_K \ldots dz_1 \geq 1 - \beta \quad (2)$$

where $f(z_1,\ldots,z_K;(\sqrt{N/2})\boldsymbol{\delta},\boldsymbol{\rho})$ represents the density of MVN with mean $(\sqrt{N/2})\boldsymbol{\delta}$ and covariance matrix $\boldsymbol{\rho}$ corresponding to $z_1,\ldots,z_K$. An iterative procedure is required to find the required sample size. The easiest way is a grid search to increase N gradually until the power under n exceeds the desired power of 1 $-\beta$, where the maximum value of the sample sizes separately calculated for each endpoint can be used as the initial values for sample size calculation.

However, this often takes much computing time. To improve the speed of the sample size calculation, Sugimoto et al. [21] and Hamasaki et al.[22] provide more efficient and practical algorithms for calculating the sample sizes. Also note that since the effect size for each co-primary endpoint is assumed to be uniform across regions, there is no difference between sample size calculations for clinical trials with co-primary endpoints conducted in multiple regions and sample size calculations for clinical trials with co-primary endpoints conducted in a single region.

Applying the results of the MRCT to a specific region

The ICH E17 says that MRCTs should investigate not only consistency in treatment effects across populations but also treatment effects in overall populations. That is, the aim of an MRCT is to show the efficacy of a drug in various global regions, and concurrently to evaluate the possibility of applying the overall trial results to each region. Suppose that we are interested in judging whether a treatment is effective in a specific region, say the sth region, where $1 \leq s \leq M$. For the kth co-primary endpoint, let D_{ik} be the observed mean difference in the ith region, D_k^{SC} be the observed mean difference from regions other than the sth region, and D_k be the observed mean difference from all regions. That is,

$$D_{ik} = (\sum_{j=1}^{N_i^T} X_{ikj})/N_i^T - (\sum_{j=1}^{N_i^C} Y_{ikj})/N_i^C,$$

$$D_k^{SC} = (\sum_{\substack{i=1 \\ i\neq s}}^{M} \sum_{j=1}^{N_i^T} X_{ikj})/(\sum_{\substack{i=1 \\ i\neq s}}^{M} N_i^T) - (\sum_{\substack{i=1 \\ i\neq s}}^{M} \sum_{j=1}^{N_i^C} Y_{ikj})/(\sum_{\substack{i=1 \\ i\neq s}}^{M} N_i^C),$$

and

$$D_k = \bar{X}_k - \bar{Y}_k.$$

Given that the overall result is significant at the α level, we establish the following criteria to judge whether the treatment is effective in the sth region:

1. $D_{s1} > \gamma_1 D_1, \ldots, D_{sK} > \gamma_K D_K$ for $0 < \gamma_i < 1$, $i = 1, \ldots, K$;
2. $D_{s1} > \gamma_1 D_1^{SC}, \ldots, D_{sK} > \gamma_K D_K^{SC}$ for $0 < \gamma_i < 1$, $i = 1, \ldots, K$;
3. $D_{s1} > h_1, \ldots, D_{sK} > h_K$ for $h_i > 0$, $i = 1, \ldots, K$.

Here, we can see that the first two criteria are to evaluate (i) whether the treatment effect in the region of interest is as large as that of the regions overall and (ii) of the other regions. Note that Criterion (i) assures that the estimated efficacy within a specific region is not smaller than a pre-specified portion of the global effect estimator.

When the sample size for the specific region is sufficiently large, the overall results will be dominated by the specific region. In this case, consistency is easier to be claimed with Criterion (i) than Criterion (ii). Therefore, Criterion (ii) tends to be more conservative than Criterion (i).

It should be noted that Criterion (i) is similar to Method 1 in the Japanese guidance. As indicated by Ikeda and Bretz [23], despite observing better results in both the entire population and the specific subpopulation, consistency sometimes can not be claimed with Method 1. This similar undesirable characteristics also exist for our Criterion (ii). Therefore, Ikeda and Bretz [23] suggested an alternative to Method 1. Let p_s denote the proportion of patients out of

$2N$ in the sth region. If we set $h_i = z_{1-\phi_i}\sigma_i\sqrt{2/Np_s}$ for given values ϕ_i, Criterion (iii) is similar to the alternative method established by Ikeda and Bretz [23]. Here ϕ_i can be thought of as the desired significant level for performing a hypothesis test for comparing the test product and the placebo control for the ith endpoint within patients from the specific region.

Sample size determination for a specific region

In the design stage, once N has been determined, special consideration should be placed on the determination of the number of subjects from the specific region in the MRCT. Per ICH E17, one important issue for conducting MRCTs is that the sample size allocation of regions should be determined such that clinically meaningful differences in treatment effects among regions can be described. Since analyses of the data from a specific region in the MRCT may not have enough statistical power, the number of subjects required for the specific region should be large enough to establish the consistency of treatment effects between the specific region and the regions overall. In this regard, ICH E17 has provided five approaches that can be considered for allocating the overall sample size to regions. Briefly, the first approach is to determine the regional sample sizes such that similar trends in treatment effects across regions can be demonstrated. The second approach is to determine the sample size needed in one or more regions such that the region-specific treatment effect preserves some pre-specified proportion of the overall treatment effect. The third approach is to enrol subjects in proportion to region size. The fourth approach is to determine the regional sample sizes so that significant results within one or more regions can be achieved. The last approach is to require a fixed minimum number of subjects in one or more regions.

In this section we suggest that, similar to the second approach suggested by ICH E17, the selected sample size should satisfy that the assurance probability of the consistency criterion in (i), (ii), or (iii), given that $\boldsymbol{\delta}$ and the overall result is significant at the α level, is maintained at a desired level, say 80%.

Let p_i denote the proportion of patients out of $2N$ in the ith region, $i = 1,\ldots,M$, where $\sum_{i=1}^{M} p_i = 1$. Also let N_i be the number of patients per group in the ith region. That is, $N_i = p_i N$. The assurance probabilities of Criteria (i)–(iii), given $\boldsymbol{\delta}$, can be represented by

$$AP_1 = P_{\boldsymbol{\delta}}(D_{s1} > \gamma_1 D_1, \ldots, D_{sK} > \gamma_K D_K | Z_1 > z_{1-\alpha}, \ldots, Z_K > z_{1-\alpha}),$$

$$AP_2 = P_{\boldsymbol{\delta}}(D_{s1} > \gamma_1 D_1^{SC}, \ldots, D_{sK} > \gamma_K D_K^{SC} | Z_1 > z_{1-\alpha}, \ldots, Z_K > z_{1-\alpha}), \quad (3)$$

and

$$AP_3 = P_{\boldsymbol{\delta}}(D_{s1} > h_1, \ldots, D_K > h_K | Z_1 > z_{1-\alpha}, \ldots, Z_K > z_{1-\alpha}),$$

Where $P_{\boldsymbol{\delta}}$ is the probability measure with respect to $\boldsymbol{\delta}$. Here we need to determine p_s to ensure that the assurance probabilities of Criteria (i)–(iii) given $\boldsymbol{\delta}$ are maintained at a desired level, say 80%. These assurance probabilities can be directly calculated by some standard normal distributions through some algebra changes; the details of the derivations of AP_1–AP_3 are given in S1 and S2 Files.

Results and discussion

Required sample sizes and assurance probabilities

Without loss of generality, we assume that we want to see whether the overall results can apply to the first region, i.e. $s = 1$. To illustrate our approach, let $K = 2$ and assume that $(\Delta_1, \Delta_2) = (3, 0.45)$

Table 1. Sample size and assurance probabilities for observing criteria (i), (ii), and (iii) given $\alpha = 0.025$, $\beta = 0.1$, $(\Delta_1, \Delta_2) = (3,0.45)$, $(\sigma_1, \sigma_2) = (6,1)$, $(\gamma_1, \gamma_2) = (0.5, 0.5)$, and $\rho_{12} = 0.1$.

p_1	N	AP_1	AP_2	AP_3	
				$\phi = 0.15$	$\phi = 0.30$
0.1	117	0.5462	0.5312	0.3276	0.5683
0.2	117	0.6786	0.6368	0.5595	0.7773
0.3	117	0.7788	0.7063	0.7294	0.8901
0.4	117	0.8568	0.7539	0.8441	0.9495
0.5	117	0.9160	0.7855	0.9171	0.9793
0.6	117	0.9578	0.8033	0.9610	0.9931
0.7	117	0.9840	0.8060	0.9854	0.9985
0.8	117	0.9967	0.7855	0.9967	0.9999
0.9	117	0.9999	0.7106	0.9999	1.0000

https://doi.org/10.1371/journal.pone.0180405.t001

and that $(\sigma_1,\sigma_2) = (6,1)$. That is, $\boldsymbol{\delta} = (0.5, 0.45)^T$. By considering $\alpha = 0.025$, $\beta = 0.1$, $(\gamma_1, \gamma_2) = (0.5, 0.5)$, and $\phi_1 = \phi_2 = \phi$, Tables 1–4 exhibit the total sample size required per group and the assurance probabilities of Criteria (i)–(iii) for $\rho_{12} = 0.1$, 0.3, 0.5, and 0.7 with various values of p_1, respectively. In Table 1, the total sample size required per group for the MRCT would be 117, which is calculated from formulas (1) and (2), for $\rho_{12} = 0.1$. The first line in Table 1 indicates that

Table 2. Sample size and assurance probabilities for observing criteria (i), (ii), and (iii) given $\alpha = 0.025$, $\beta = 0.1$, $(\Delta_1, \Delta_2) = (3,0.45)$, $(\sigma_1, \sigma_2) = (6,1)$, $(\gamma_1, \gamma_2) = (0.5, 0.5)$, and $\rho_{12} = 0.3$.

p_1	N	AP_1	AP_2	AP_3	
				$\phi = 0.15$	$\phi = 0.30$
0.1	115	0.5690	0.5549	0.3577	0.5891
0.2	115	0.6937	0.6545	0.5798	0.7861
0.3	115	0.7880	0.7198	0.7403	0.8932
0.4	115	0.8617	0.7646	0.8489	0.9502
0.5	115	0.9180	0.7944	0.9191	0.9795
0.6	115	0.9583	0.8112	0.9613	0.9931
0.7	115	0.9842	0.8135	0.9852	0.9985
0.8	115	0.9967	0.7944	0.9967	0.9999
0.9	115	0.9999	0.7238	0.9999	1.0000

https://doi.org/10.1371/journal.pone.0180405.t002

Table 3. Sample size and assurance probabilities for observing criteria (i), (ii), and (iii) given $\alpha = 0.025$, $\beta = 0.1$, $(\Delta_1, \Delta_2) = (3,0.45)$, $(\sigma_1, \sigma_2) = (6,1)$, $(\gamma_1, \gamma_2) = (0.5, 0.5)$, and $\rho_{12} = 0.5$.

p_1	N	AP_1	AP_2	AP_3	
				$\phi = 0.15$	$\phi = 0.30$
0.1	114	0.5955	0.5821	0.3915	0.6142
0.2	114	0.7128	0.6761	0.6048	0.7988
0.3	114	0.8009	0.7372	0.7559	0.8991
0.4	114	0.8697	0.7790	0.8574	0.9525
0.5	114	0.9223	0.8068	0.9232	0.9804
0.6	114	0.9603	0.8224	0.9633	0.9934
0.7	114	0.9848	0.8249	0.9859	0.9985
0.8	114	0.9968	0.8069	0.9970	0.9999
0.9	114	0.9999	0.7410	0.9999	1.0000

https://doi.org/10.1371/journal.pone.0180405.t003

Table 4. Sample size and assurance probabilities for observing criteria (i), (ii), and (iii) given $\alpha = 0.025$, $\beta = 0.1$, $(\Delta_1, \Delta_2) = (3, 0.45)$, $(\sigma_1, \sigma_2) = (6, 1)$, $(\gamma_1, \gamma_2) = (0.5, 0.5)$, and $\rho_{12} = 0.7$.

p_1	N	AP_1	AP_2	AP_3	
				$\phi = 0.15$	$\phi = 0.30$
0.1	111	0.6250	0.6125	0.4266	0.6408
0.2	111	0.7341	0.7001	0.6303	0.8123
0.3	111	0.8154	0.7566	0.7709	0.9050
0.4	111	0.8785	0.7951	0.8653	0.9551
0.5	111	0.9271	0.8206	0.9268	0.9814
0.6	111	0.9621	0.8353	0.9649	0.9936
0.7	111	0.9853	0.8370	0.9865	0.9983
0.8	111	0.9966	0.8206	0.9973	0.9999
0.9	111	0.9999	0.7603	0.9999	1.0000

https://doi.org/10.1371/journal.pone.0180405.t004

if the proportion of patients out of the total number of patients in the study is 0.10, the assurance probabilities of Criteria (i) and (ii) are respectively 0.55, 0.53, while the assurance probabilities for criteria (iii) with corresponding to $\phi = 0.15$ and $\phi = 0.30$ are respectively 0.33 and 0.57. From Table 1, to achieve assurance probability at the 80% level, the sample size for the first region has to be around 40% of the overall sample size for criteria (i), and to be around 60% for criterion (ii). On the other hand, the assurance probabilities of Criterion (iii) will reach 80% when the values of p_1 are 40% and 30% for $\phi = 0.15$ and $\phi = 0.30$ respectively. Note that the sample size required per group is the minimum N satisfying (1) and (2); therefore, the assurance probabilities for criteria (i), (ii), and (iii) must increase more if the sample size in a practical trial is larger than N.

In Tables 1–4, we see the following phenomena. First of all, we found that as p_1 increases, the assurance probability of Criterion (i) increases. This is due to the fact that as p_1 increases, the observed overall results D_k's will be increasingly dominated by the observed result from the first region, D_{1k}'s. Secondly, we have also observed that as p_1 increases, the assurance probability of Criterion (ii) increases first and then decrease later. This occurs because the observed result from regions other than the first region, D_k^{SC}'s, is gradually dominated by D_{1k}'s at first and is then completely dominated by D_{1k}'s later as p_1 increases. Also, the assurance probability of Criterion (iii) increases when p_1 increases, since the h_i's decrease as p_1 increases.

Another feature we observed is that $AP_1 > AP_2$ in Tables 1, 2, 3 and 4, given p_1 and γ_k. This is due to the fact that

$$\begin{aligned}
&P_{\boldsymbol{\delta}}(D_{1k} > \gamma_k D_k^{SC},\ k = 1, \ldots, K | Z_k > z_{1-\alpha},\ k = 1, \ldots, K) \\
&= P_{\boldsymbol{\delta}}\left(D_{1k} > \frac{\gamma_k}{1 - p_1 + \gamma_k p_1} D_k,\ k = 1, \ldots, K | Z_k > z_{1-\alpha},\ k = 1, \ldots, K\right) \\
&< P_{\boldsymbol{\delta}}(D_{1k} > \gamma_k D_k,\ k = 1, \ldots, K | Z_k > z_{1-\alpha},\ k = 1, \ldots, K)
\end{aligned}$$

since

$$\begin{aligned}
&0 < 1 - p_1 + \gamma_k p_1 \\
&= 1 - (1 - \gamma_k) p_1 \\
&< 1.
\end{aligned}$$

Like Ikeda and Bretz [23] suggested, for Criterion (iii), it may be able to link the choice of ϕ to $(\gamma_1, \gamma_2) = (0.5, 0.5)$ in order to ensure the same level of strictness of Criterion (i). For

example, in Table 1, setting $\phi = 0.17$ in Criterion (iii) would closely ensure a similar level of strictness as Criterion (i) with $p_1 = 0.4$. Another point we wish to make is that the assurance probabilities of all criteria increase as ρ_{12} increases. This makes intuitive sense because these two co-primary endpoints look more alike.

Numerical example

In this section, we provide an example to illustrate a practical application of our method. A randomized, double-blind, active-controlled MRCT will be conducted in patients with mild to moderate AD for comparing a new treatment and a placebo control. In this trial, patients age 50 or older with a diagnosis of uncomplicated AD are planned to be recruited from three regions: Taiwan, the European Union, and the United States. The primary endpoints are the change from baseline of ADAS-cog at week 24 and the CIBIC plus value at week 24. Based on the results observed in a previous exploratory study, the differences of change in ADAS-cog score from the baseline and the CIBIC-plus at week 24 between the test drug and placebo are expected to be 2.88 and 0.44, respectively. Also the standard deviations for both groups for change in ADAS-cog score from the baseline and the CIBIC-plus at week 24 are respectively equal and are assumed to be 6.15 and 0.92. With $\rho_{12} = 0, 0.3, 0.5, 0.8$, $\alpha = 0.025$, and $\beta = 0.1$, the sample sizes required per group determined by (1) and (2) are as follows:

$$n = \begin{cases} 116 \text{ if } \rho_{12} = 0, \\ 114 \text{ if } \rho_{12} = 0.3, \\ 112 \text{ if } \rho_{12} = 0.5, \\ 107 \text{ if } \rho_{12} = 0.8. \end{cases}$$

In addition, in order to demonstrate an overall treatment effect from all regions the sponsor is also interested in assessing whether the overall results from the multi-regional trial can be bridged to Taiwan if the overall treatment effect shows statistical significance. In this regard, the proportion of the patients recruited in Taiwan needs to be determined during the design phase of the trial to preserve the probability of establishing consistency between Taiwan and all other regions. Suppose that similarity criterion (i) is used, and that $\gamma_1 = 0.5$ and $\gamma_2 = 0.5$ are chosen. To insure the assurance probability of AP_1 at the 80% level, the sample sizes required per group, n_S, from Taiwan patients with respect to $\rho_{12} = 0, 0.3, 0.5$, and 0.8 are shown below:

$$n_S = \begin{cases} 116 \times 33\% \approx 39 \text{ or } 40 \text{ if } \rho_{12} = 0, \\ 114 \times 32\% \approx 37 \text{ or } 38 \text{ if } \rho_{12} = 0.3, \\ 112 \times 30\% \approx 34 \quad\quad \text{if } \rho_{12} = 0.5, \\ 107 \times 27\% \approx 29 \text{ or } 30 \text{ if } \rho_{12} = 0.8. \end{cases}$$

Conclusions

The aim of an MRCT is to show the efficacy of a drug in various global regions, and simultaneously to evaluate the possibility of applying the overall trial results to each region. However, in MRCTs sponsors are challenged by how to demonstrate consistency between a specific region and the overall results. In this paper, three criteria have been established to assess the similarity between a specific region and the overall regions in an MRCT with multiple co-primary endpoints. Regulators and sponsors can easily adopt these criteria to conduct statistical assessments of the consistency of treatment effects between the specific region and the entire trial, and consequently to help registration of the new drug in the specific region.

On the other hand, the 11th Q&A for ICH E5 states, "It may be desirable in certain situations to achieve the goal of bridging by conducting a multi-regional trial under a common protocol that includes sufficient numbers of patients from each of multiple regions to reach a conclusion about the effect of the drug in all regions." Therefore, the sample size determination for each region is another challenge for regulators and sponsors. With the three criteria we established, the sample size required for a specific region can easily be determined so that there is a high probability of observing a consistent trend in treatment effect between the specific region and the entire MRCT. In this paper, we do not particularly recommend any criterion for evaluating the consistency of treatment effects between the entire region and the specific region.

Although our approach is easy to use, the selection of the magnitude γ_i's consistency trend raises an important issue. In this regard, the Japanese guidance suggests that the magnitude be 0.5 or greater for the first criterion when the number of primary endpoints for the MRCT is only one. Our suggestion is that the determination of γ_i should be discussed between the regulatory agency in the specific region and the trial sponsor. Most importantly, all differences in race, diet, environment, culture, and medical practice among regions should be considered.

It should be noted that, in our approach, the sample size calculation for the specific region did not have a closed-form expression. For conducting an MRCT with only one primary endpoint, Ikeda and Bretz [23] discussed the methods proposed in the Japanese regulatory guidance document and derived closed-form expressions for the resulting probabilities, which required the evaluation of multivariate normal or t probabilities between the overall effect and the effect in Japan. In addition, they proposed a different method of calculating the probability of observing a consistent trend based on Method 1 in the Japanese regulatory guidance. Ikeda and Bretz's work is worthy of being extended to the MRCT with multiple co-primary endpoints.

When more than one primary endpoint is viewed as important in a clinical trial, a decision must be made as to whether it is desirable to evaluate the joint effects on at least one or even all of the endpoints. This decision defines the alternative hypothesis to be tested and provides a framework for trial design. This article discusses only the former situation, where a trial is designed to evaluate the joint effects of a new treatment compared to any control treatment on all of the primary endpoints as seen in AD clinical trials. On the other hand, the latter situation—i.e., designing the trial to evaluate an effect on at least one of the primary endpoints is referred to as "multiple primary endpoints" (Offen et al. [15])—and many methods for dealing with such multiple primary endpoints have been proposed (e.g., see the extensive references in Dmitrienko et al [24]). Similarly, as in multiple co-primary endpoints, the power for detecting an effect on at least one endpoint—which is called "disjunctive power" (Senn and Bretz [19]) or "minimal power" (Westfall et al. [20]))—can be defined and extended.

Another issue we want to point out is that in this paper, it is assumed that the outcome variances are known for the sample size calculation. In actual practice, the outcome variances are not known and should be estimated from some data. In fact, extensive literature of results of similar trials may exist, and thus the variability associated with the primary endpoints can also be found in literature. For methods for unknown variance, the major change is that the power function will be evaluated based on a non-central multivariate t-distribution. For clinical trials with multiple co-primary endpoints, Sozu et al. [18] discussed a method for the unknown variance case and showed that the calculated sample size is nearly equivalent to that for the known variance in the setting of 80% or 90% power at 2.5% significance level for one-sided test. They showed that the sample size per group calculated using the method based on the unknown variance needs generally one more subject than that using the method based on the known variance. This is a very similar result observed as in a single primary endpoint case. Therefore,

sample size calculation based on a known variance provides a reasonable approximation for the unknown variances case.

Similarly, the correlation is usually unknown and thus must be estimated by (1) using data from pilot studies or proceeding clinical trials (e.g., Phase II trials), or by (2) borrowing information from external existing data when incorporating correlation into sample size calculation. In some disease areas, the correlations among the endpoints have been known. For example, Offen et al. [15] provides a list of known disease are as that the regulatory agency requires for co-primary endpoints when evaluating the effects of a new treatment; the list includes possible correlations among endpoints for each disease area.

The proposed criteria can be extended from one to multiple regions. For example, after the MRCT has demonstrated a statistically significant overall treatment effect, we can bridge the results of the MRCT to all regions if

$$D_{i1} > \gamma_k^i D_1, \ldots, D_{iK} > \gamma_K^i D_K \text{ for } 0 < \gamma_k^i < 1,\ i = 1, \ldots, M,\ k = 1, \ldots,\ K.$$

Here γ_k^i represents the threshold of consistency trend for the kth endpoint in the $i$thregion. Our research work here assumes that the effect size for each co-primary endpoint and the correlations among endpoints are both uniform across regions. Since MRCTs recruit subjects from many countries around the world, it might be expected that there is a difference in treatment effect or in correlations among endpoints due to regional difference (e.g., ethnic difference).Thus, the sample size calculation for MRCTs based on the assumption that the effect size for each co-primary endpoint and the correlations among endpoints are uniform across regions might be impractical. Future work is being pursued to address this issue.

Supporting information

S1 File. Derivations of the three assurance probabilities AP_1, AP_2, and AP_3.
(PDF)

S2 File. Codes of R software for AP_1, AP_2, and AP_3.
(PDF)

Acknowledgments

The work in this paper was supported by a grant from the Ministry of Science and Technology, Taiwan (MOST103-2118-M-400-004).Thanks are due to two referees for his or her detailed, constructive, and thoughtful comments and suggestions, which we believe have led to significant improvements to this paper.

Author Contributions

Conceptualization: Wong-Shian Huang, Hui-Nien Hung, Toshimitsu Hamasaki, Chin-Fu Hsiao.

Formal analysis: Wong-Shian Huang.

Funding acquisition: Chin-Fu Hsiao.

Investigation: Wong-Shian Huang, Hui-Nien Hung, Toshimitsu Hamasaki, Chin-Fu Hsiao.

Methodology: Wong-Shian Huang, Hui-Nien Hung, Toshimitsu Hamasaki, Chin-Fu Hsiao.

Project administration: Chin-Fu Hsiao.

Software: Wong-Shian Huang.

Supervision: Chin-Fu Hsiao.

Validation: Hui-Nien Hung, Toshimitsu Hamasaki, Chin-Fu Hsiao.

Visualization: Wong-Shian Huang, Chin-Fu Hsiao.

Writing – original draft: Wong-Shian Huang, Hui-Nien Hung, Toshimitsu Hamasaki, Chin-Fu Hsiao.

Writing – review & editing: Wong-Shian Huang, Hui-Nien Hung, Toshimitsu Hamasaki, Chin-Fu Hsiao.

References

1. International Conference on Harmonization. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Q&A for the ICH E5 Guideline on Ethnic Factors in the Acceptability of Foreign Data. 2006. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E5_R1/Q_As/E5_Q_As__R5_.pdf. Accessed on May 3, 2017.
2. International Conference on Harmonization. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH E17 Guideline on General principle on planning/designing Multi-Regional Clinical Trials. 2016. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E17/E17_Step2.pdf. Accessed on May 3, 2017.
3. Ministry of Health, Labour and Welfare of Japan. Basic Principles on Global Clinical Trials. 2007. http://www.pmda.go.jp/files/000153265.pdf. Accessed on May 3, 2017.
4. Quan H, Zhao PL, Zhang J, Roessner M, Aizawa K. Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance. Pharmaceutical Statistics. 2010; 9(2):100–112. https://doi.org/10.1002/pst.380 PMID: 19499510
5. Kawai N, Chuang-Stein C, Komiyama O, Li Y. An approach to rationalize partitioning sample size into individual regions in a multiregional trial. Drug Information Journal. 2008; 42(2):139–147.
6. Tsou HH, James Hung HM, Chen YM, Huang WS, Chang WJ, Hsiao CF. Establishing consistency across all regions in a multi-regional clinical trial. Pharmaceutical Statistics. 2012; 11(4): 295–299. https://doi.org/10.1002/pst.1512 PMID: 22504851
7. Uesaka H. Sample size allocation to regions in a multiregional trial. Journal of Biopharmaceutical Statistics. 2009; 19(4): 580–594. https://doi.org/10.1080/10543400902963185 PMID: 20183427
8. Ko FS, Tsou HH, Liu JP, Hsiao CF. Sample size determination for a specific region in a multiregional trial. Journal of Biopharmaceutical Statistics. 2010; 20(4): 870–885. https://doi.org/10.1080/10543401003618900 PMID: 20496211
9. Tsou HH, Chien TY, Liu JP, Hsiao CF. A consistency approach to evaluation of bridging studies and multi-regional trials. Statistics in Medicine. 2011; 30(17): 2171–2186. https://doi.org/10.1002/sim.4251 PMID: 21590701
10. Chen CT, Hung HMJ, Hsiao CF. Design and evaluation of multiregional trials with heterogeneous treatment effect across regions. Journal of Biopharmaceutical Statistics. 2012; 22(5): 1037–1050. https://doi.org/10.1080/10543406.2012.701585 PMID: 22946948
11. Huang Y, Chang WJ, Hsiao CF. An empirical Bayes approach to evaluation of results for a specific region in multiregional clinical trials. Pharmaceutical Statistics. 2013; 12(2): 59–64. https://doi.org/10.1002/pst.1553 PMID: 23319408
12. Committee for Medicinal Products for Human Use (CHMP). Guideline on medicinal products for the treatment of Alzheimer's disease and other dementias (CPMP/EWP/553/95 Rev.1). European Medical Agency, London, UK. 2008. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003562.pdf. Accessed on May 3, 2017.
13. Food and Drug Administration. Guidance for Industry. Alzheimer's disease: developing drugs for the treatment of early stage disease. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD, USA. 2013. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm338287.pdf. Accessed on May 3, 2017.
14. Hung HM, Wang SJ. Some controversial multiple testing problems in regulatory applications. Journal of Biopharmaceutical Statistics. 2009; 19:1–11. https://doi.org/10.1080/10543400802541693 PMID: 19127460
15. Offen W, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, et al. Multiple co-primary endpoints: medical and statistical solutions. Drug Information Journal. 2007; 41:31–46.

16. Sozu T, Sugimoto T, Hamasaki T, Evans SR. Sample size determination in clinical trials with multiple endpoints. Cham: Springer International Publishing; 2015.
17. Sozu T, Kanou T, Hamada C, Yoshimura I. Power and sample size calculations in clinical trials with multiple primary variables. Japanese Journal of Biometrics. 2006; 27:83–96.
18. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. Journal of Biopharmaceutical Statistics.2011; 21:650–668. https://doi.org/10.1080/10543406.2011.551329 PMID: 21516562
19. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. Pharmaceutical Statistics. 2007; 161–170. https://doi.org/10.1002/pst.301 PMID: 17674404
20. Westfall PH, Tobias RD, Wolfinger RD. Multiple comparisons and multiple tests using SAS, 2nd edition, 2011. Cary, NC: SAS Institute Inc.
21. Sugimoto T, Sozu T, Hamasaki T. A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. Pharmaceutical Statistics. 2012; 11:118–128. https://doi.org/10.1002/pst.505 PMID: 22415870
22. Hamasaki T, Sugimoto T, Evans SR, Sozu T. Sample size determination for clinical trials with co-primary outcomes: exponential event-times. Pharmaceutical Statistics. 2013; 12:28–34. https://doi.org/10.1002/pst.1545 PMID: 23081932
23. Ikeda K, Bretz F. Sample size and proportion of Japanese patients in multi-regional trials. Pharmaceutical Statistics. 2010; 207–216. https://doi.org/10.1002/pst.455 PMID: 20872621
24. Dmitrienko A, Tamhane AC, Bretz F. Multiple testing problems in pharmaceutical statistics. 2010, Chapman & Hall/CRC, Boca Raton