Statistical implications of extrapolating the overall result to the target region in multi-regional clinical trials

Seung-Ho Kang[1,a], Saemina Kim[a]

[a]Department of Applied Statistics, Yonsei University, Korea

Abstract

The one of the principles described in ICH E9 is that only results obtained from pre-specified statistical methods in a protocol are regarded as confirmatory evidence. However, in multi-regional clinical trials, even when results obtained from pre-specified statistical methods in protocol are significant, it does not guarantee that the test treatment is approved by regional regulatory agencies. In other words, there is no so-called global approval, and each regional regulatory agency makes its own decision in the face of the same set of data from a multi-regional clinical trial. Under this situation, there are two natural methods a regional regulatory agency can use to estimate the treatment effect in a particular region. The first method is to use the overall treatment estimate, which is to extrapolate the overall result to the region of interest. The second method is to use regional treatment estimate. If the treatment effect is completely identical across all regions, it is obvious that the overall treatment estimator is more efficient than the regional treatment estimator. However, it is not possible to confirm statistically that the treatment effect is completely identical in all regions. Furthermore, some magnitude of regional differences within the range of clinical relevance may naturally exist for various reasons due to, for instance, intrinsic and extrinsic factors. Nevertheless, if the magnitude of regional differences is relatively small, a conventional method to estimate the treatment effect in the region of interest is to extrapolate the overall result to that region. The purpose of this paper is to investigate the effects produced by this type of extrapolation via estimations, followed by hypothesis testing of the treatment effect in the region of interest. This paper is written from the viewpoint of regional regulatory agencies.

Keywords: consistency, type I error rate, power, intrinsic factor, extrinsic factor

1. Multi-regional clinical trials from the viewpoint of regional regulatory agencies

Currently, multi-regional clinical trials (MRCTs) play a major role in providing evidence of the efficacy and safety of new drugs. MRCTs are studies planned with the objective of the world-scale development and approval of new drugs. In MRCTs, study sites at multiple countries and regions participate in a single study based on a common protocol and the trials are conducted simultaneously (ICH, 2006, 2017; MHLW, 2007). In the planning stages of a MRCT, an important premise is that there can be no, or reasonably small, regional variation. This premise of MRCTs is usually supported by biological and medical knowledge about the disease; the mechanism of action of the compound; the medical practice characteristics in the individual regions (e.g., treatment guidelines and the availability of concomitant therapies); the health systems, and the available nonclinical, preclinical, and early clinical data. Usually, a small difference in the magnitude of the treatment effect is acceptable.

If a substantial regional variation is expected, a global development strategy using MRCTs may not be appropriate (Chen *et al.*, 2011).

After a MRCT is finished, there are two main objectives in a conventional analysis of the MRCT. The primary is to demonstrate the efficacy of a drug in all participating regions. The second is, after the overall effect is shown, to evaluate the possibility of applying the overall trial results to each region by checking the consistency of the treatment effect across regions. Hence, an assessment of the consistency of the treatment effect across regions is a key issue in relation to MRCTs (Chen *et al.*, 2010; Liu *et al.*, 2013; Quan *et al.*, 2010a, 2013; Tsou *et al.*, 2012; ICH, 2017). Furthermore, it is important to assign a sufficient number of patients to each region in order to evaluate the consistency of the treatment effect across regions (Kawai *et al.*, 2008; Ko *et al.*, 2010; Ikeda and Bretz, 2010; Quan *et al.*, 2010b; Uesaka, 2009; ICH, 2017).

In the analysis of data from MRCTs there is a very important thing to keep in mind, which there is no a so-called "global approval" of a medical product. From a regulatory standpoint, an approval decision is regional; that is, each regional regulatory agency makes its own decision given an identical set of data from a MRCT. This means that the same set of data from a MRCT may be analyzed differently to obtain approval from different regional regulatory agencies. The challenge is how to utilize the entire dataset to assess individual region effects, and how to leverage the data from outside the local region in a MRCT to make approval decisions in the local region. Therefore, it is very important to analyze MRCTs data from the viewpoint of regional regulatory agencies (Chen *et al.*, 2012a; Tsong *et al.*, 2012; Tsou *et al.*, 2010).

2. Motivation

For convenience, let the target region denote the region of interest. Let the treatment effects in both the target region and the non-target region be denoted by δ_T and δ_{NT}, respectively. It is assumed that the MRCT has already been conducted. From the viewpoint of the target regulatory agency, an important issue is to make statistical inference for the treatment effect in the target region (δ_T). Keeping this fact in mind, we consider the following three cases.

2.1. $|\delta_T - \delta_{NT}| = 0$

In this case, the treatment effects in both the target region and the non-target region are completely identical. The target regulatory agency can make accurate statistical inference for the treatment effect in the target region (δ_T) based on the information from both the target region and the non-target region. The analysis strategy of MRCTs described in Section 1 is based on the assumption of $|\delta_T - \delta_{NT}| = 0$. In other words, when $|\delta_T - \delta_{NT}| = 0$, the target regulatory agency can apply the overall trial results to its own region without any problems. Patients in the non-target region are considered to be identical to those in the target region for the purpose of estimating the treatment effect in the target region (δ_T). In extreme cases, the target regulatory agency is still able to make accurate statistical inference for the treatment effect in the target region (δ_T) even when there are no patients from the target region, because $\delta_T = \delta_{NT}$.

However, the question arises of how we can be sure that the treatment effects in both the target region and the non-target region are completely identical ($\delta_T = \delta_{NT}$)? It appears to be reasonable to assume that $\delta_T \neq \delta_{NT}$ in MRCTs due to various intrinsic and extrinsic factors, although the differences in treatment effects between the two regions can be either small or substantial (ICH, 2017).

If $\delta_T \neq \delta_{NT}$, can the target regulatory agency borrow information from the non-target region? In a mathematical sense, the answer appears to be no, because the treatment effect in the target region

differs from that in the non-target region. However, if the difference is small and clinically not meaningful, the answer may be yes. Therefore, the key issue is how different the treatment effect in the target region is from that in the non-target region when $\delta_T \neq \delta_{NT}$.

2.2. $|\delta_T - \delta_{NT}| \neq 0$, but the difference is substantial

In this case, a global development strategy using MRCTs may not be appropriate, and conducting separate clinical trials in each region may be a better strategy. If the fact that the difference in the treatment effect between the target region and the non-target region is substantial comes to light after the MRCT is conducted, it is obvious that the target regulatory agency cannot borrow information from the non-target region. In such a case, an estimate of the treatment effect only based on the target region is a reliable estimate of the treatment effect in the target region (δ_T). In this paper, we will not discuss this case anymore.

2.3. $|\delta_T - \delta_{NT}| \neq 0$, but the difference is small and clinically not meaningful

In this case, it may be possible for the target regulatory agency to borrow information from the non-target region, as the difference in the treatment effect between the target region and the non-target region is small and clinically not meaningful. Hence, the overall treatment estimator may be more reliable than the regional treatment estimator, i.e., the treatment effect estimator only based on the target region.

The purpose of this paper is to provide a statistical rationale which applies when the overall treatment estimator is better than the regional treatment estimator in order to assess the treatment effect in the target region (δ_T). In this paper, for convenience it is assumed that the sizes of the treatment effect across the non-target regions are the same. And we consider only the fixed-effect model. Consideration of the random-effect model is discussed in Section 7.

3. Notation and review

This paper considers only the viewpoint of the target regulatory agency, as approval decisions are regional. We consider parallel designs for comparing a test product and a placebo. Let X_{ij} and Y_{ij} be the primary endpoints, with large values representing better efficacy for j^{th} patients receiving the test product and the placebo control, respectively, in the ith region, where $i = T, NT$ denote the target and the non-target region, respectively. It is also assumed that the number of patients in the test and placebo groups in each region are identical, that is, $j = 1, 2, \ldots, n_i$ for $i = T, NT$.

We also assume the following:

$$X_{ij} \sim N\left(\mu_{iX}, \sigma^2\right), \qquad Y_{ij} \sim N\left(\mu_{iY}, \sigma^2\right).$$

The primary hypothesis for testing the overall treatment effect is given by

$$H_0 : \delta \leq 0 \quad \text{versus} \quad H_A : \delta > 0, \tag{3.1}$$

where δ denotes the overall treatment effect. The definition is provided in (3.3) below after the overall treatment estimator ($\hat{\delta}$) is introduced. The primary alternative hypothesis in (3.1) states that the test product is effective at a global level.

However, the target regulatory agency is more interested in the regional treatment effect in the target region; the hypotheses of interest are expressed as

$$H_{T0} : \delta_T \leq 0 \quad \text{versus} \quad H_{TA} : \delta_T > 0, \tag{3.2}$$

where $\delta_T = \mu_{TX} - \mu_{TY}$. The alternative hypothesis in (3.2) states that the test product is effective in the target region. If we consider the regional treatment effect as a fixed effect in a fixed-effect model, letting $\delta_i = \mu_{iX} - \mu_{iY}$ be the treatment effect of region i ($i = T, NT$), the regional treatment estimator for region i is

$$\hat{\delta}_i = \bar{X}_i - \bar{Y}_i \sim N\left(\delta_i, \frac{2}{n_i}\sigma^2\right),$$

where

$$\bar{X}_T = \frac{1}{n_T}\sum_{j=1}^{n_T} X_{T,j}, \qquad \bar{Y}_T = \frac{1}{n_T}\sum_{j=1}^{n_T} Y_{T,j},$$

$$\bar{X}_{NT} = \frac{1}{n_{NT}}\sum_{j=1}^{n_{NT}} X_{NT,j}, \qquad \bar{Y}_{NT} = \frac{1}{n_{NT}}\sum_{j=1}^{n_{NT}} Y_{NT,j}.$$

A commonly used overall treatment estimator ($\hat{\delta}$) is a weighted average of the regional treatment estimators:

$$\hat{\delta} = \frac{n_T\hat{\delta}_T + n_{NT}\hat{\delta}_{NT}}{N} = p\hat{\delta}_T + (1-p)\hat{\delta}_{NT},$$

where $N = n_T + n_{NT}$ is the total number of patients in the trial in each treatment arm and $p = n_T/N$. In the fixed-effect model, the overall treatment effect (δ) is usually defined as follows (Quan *et al.*, 2013, 2014).

$$\delta = p\delta_T + (1-p)\delta_{NT}. \tag{3.3}$$

The regional treatment estimator ($\hat{\delta}_T$) uses data from only the target region. It ignores data from the non-target region. Owing to the reduced sample size for the target region, it has considerable variability and may not be an efficient estimator of the treatment effect for the target region. When there is a significant overall treatment effect, we may use the overall treatment estimator as a more precise estimator of the treatment effect in the target region.

As discussed in Section 2.1, if $\delta_T = \delta_{NT}$, it is clear that the target regulatory agency can use the overall treatment estimator ($\hat{\delta}$) as an estimate of the treatment effect in the target region (δ_T). Moreover, there are no issues to discuss. Therefore, in this paper, it is assumed that the treatment effects in both the target region and the non-target region are not completely identical ($\delta_T \neq \delta_{NT}$), but the difference is small.

In such cases, it is well-known that the power of interaction test to detect treatment-by-region interaction is low. With the help of SAS PROC GLMPOWER, the power of interaction test is plotted in Figure 1 when $N = 1,000$, $\delta_{NT} = 0.3$, and $p = 0.2$ as the value of δ_T changes. When $\delta_T \neq \delta_{NT}$ and the difference between δ_T and δ_{NT} is small, the power of interaction test is low. For example, when $\delta_T = 0.1$ and $\delta_{NT} = 0.2$, the power is less than 10%. Therefore, the conclusion that the treatment effect is consistent across regions is very likely, despite the small difference in the treatment effect.

In Section 4, we compare $\hat{\delta}_T$ and $\hat{\delta}$ in terms of the estimation of the treatment effect in the target region (δ_T). In Sections 5 and 6, we investigate the type I error rate and power when the overall treatment effect estimator ($\hat{\delta}$) is employed to test the treatment effect in the target region (δ_T).
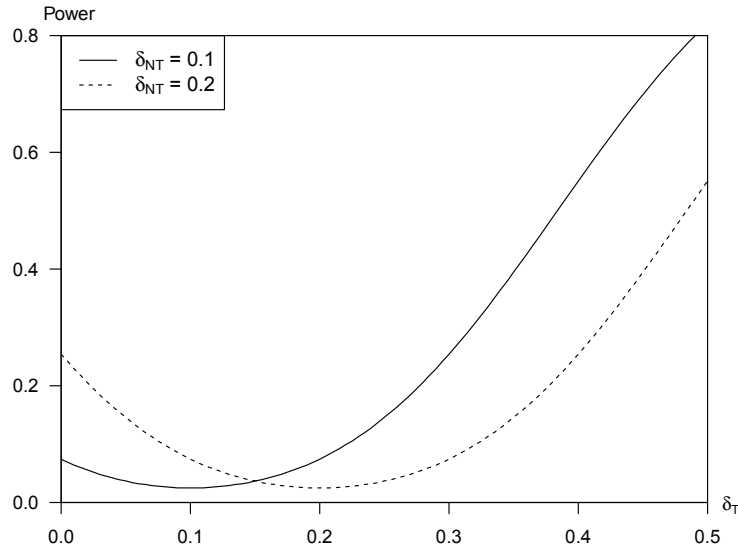
Figure 1: *The power of the interaction test when N = 1000 and p = 0.2.*

4. The comparison of $\hat{\delta}_T$ and $\hat{\delta}$ in terms of the mean square error

In this section, we compare the performance capabilities of the two estimators ($\hat{\delta}_T$ and $\hat{\delta}$) in order to estimate the treatment effect in the target region (δ_T) when the treatment effects in both the target region and the non-target region differ slightly ($\delta_T \neq \delta_{NT}$). The key idea is that the mean square error (MSE) of $\hat{\delta}$ is smaller than that of $\hat{\delta}_T$, although $\hat{\delta}$ is not an unbiased estimator of δ_T. This is true because the sample size in $\hat{\delta}$ is usually much larger than that in $\hat{\delta}_T$, and the difference between the treatment effects in the target region and the non-target region ($|\delta_T - \delta_{NT}|$) is assumed to be small.

4.1. Unbiasedness

It is clear that $E[\hat{\delta}_T] = \delta_T$; hence, the regional treatment estimator ($\hat{\delta}_T$) is unbiased. However, when $|\delta_T - \delta_{NT}| \neq 0$, because

$$E\left[\hat{\delta}\right] = E\left[p\hat{\delta}_T + (1-p)\hat{\delta}_{NT}\right] = p\delta_T + (1-p)\delta_{NT} \neq \delta_T,$$

the overall treatment estimator ($\hat{\delta}$) is not an unbiased estimator of δ_T.

4.2. The mean square error

Given that the regional treatment effect estimator ($\hat{\delta}_T$) follows the following distribution

$$\hat{\delta}_T = \bar{X}_T - \bar{Y}_T \sim N\left(\delta_T, \frac{2}{n_T}\sigma^2\right),$$

the MSE of $\hat{\delta}_T$ is given as follows.

$$\text{MSE}\left[\hat{\delta}_T\right] = \text{Bias}^2 + \text{Variance} = \frac{2}{n_T}\sigma^2.$$

(a) $N = 400$

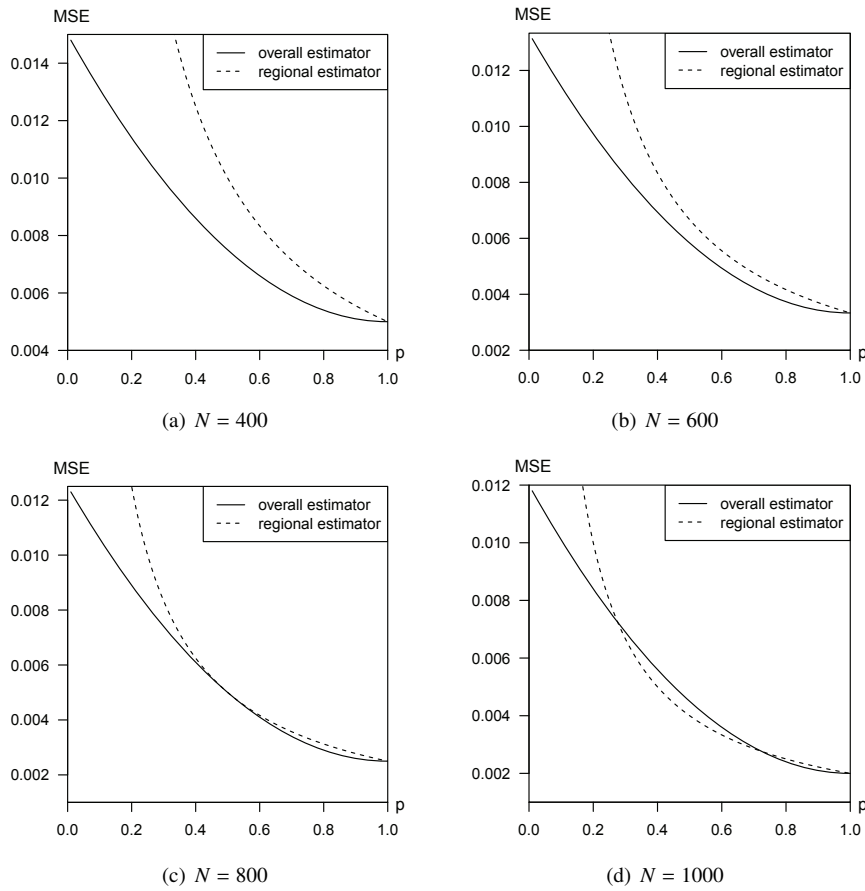(b) $N = 600$

(c) $N = 800$

(d) $N = 1000$

Figure 2: *The mean square errors (MSEs) of the overall treatment effect estimator and the regional treatment effect estimator when* $|\delta_T - \delta_{NT}| = 0.1$.

Similarly, because

$$p^2 \frac{2}{n_T} \sigma^2 + (1 - p)^2 \frac{2}{n_{NT}} \sigma^2 = \frac{2}{N} \sigma^2,$$

we have

$$\hat{\delta} = p\hat{\delta}_T + (1 - p)\hat{\delta}_{NT} \sim N\left(p\delta_T + (1 - p)\delta_{NT}, \frac{2}{N}\sigma^2\right).$$

Therefore, the MSE of $\hat{\delta}$ is given as follows.

$$\text{MSE}\left[\hat{\delta}\right] = \left[(1 - p)(\delta_{NT} - \delta_T)\right]^2 + \frac{2}{N}\sigma^2.$$

When $|\delta_T - \delta_{NT}| = 0$, it is clear that the MSE of $\hat{\delta}$ is smaller than that of $\hat{\delta}_T$. When the difference between the treatment effects in the target region and the non-target region ($|\delta_T - \delta_{NT}|$) is assumed to
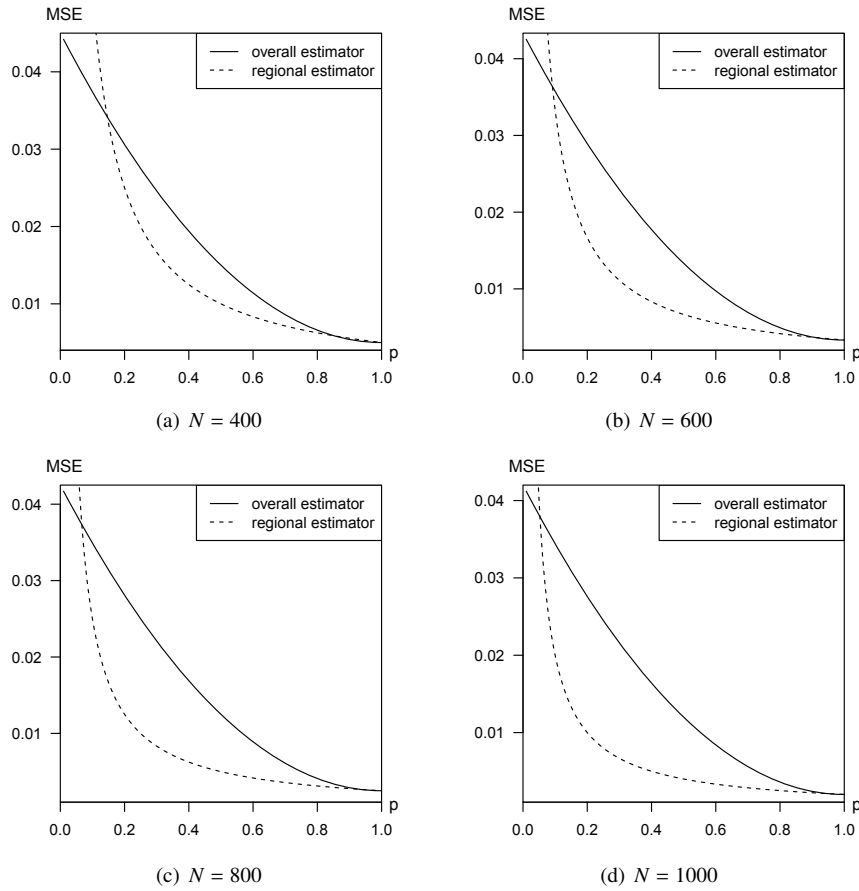
(a) $N = 400$

(b) $N = 600$

(c) $N = 800$

(d) $N = 1000$

Figure 3: *The mean square errors (MSEs) of the overall treatment effect estimator and the regional treatment effect estimator when $|\delta_T - \delta_{NT}| = 0.2$.*

exist but is small, Figures 2 and 3 are helpful when attempting to understand the relationship between the MSEs of $\hat{\delta}$ and $\hat{\delta}_T$.

If we compare the two MSEs analytically, we have

$$\text{MSE}\left[\hat{\delta}\right] = \left[(1 - p)(\delta_{NT} - \delta_T)\right]^2 + \frac{2}{N}\sigma^2 < \frac{2}{Np}\sigma^2 = \text{MSE}\left[\hat{\delta}_T\right]$$

$$\Leftrightarrow |\delta_{NT} - \delta_T| < \sqrt{\frac{2}{N}\frac{1}{p(1-p)}\sigma^2}. \tag{4.1}$$

The inequality in (4.1) shows us the maximum allowable difference of the treatment effect between the target region and the non-target region ($|\delta_{NT} - \delta_T|$) such that the MSE of $\hat{\delta}$ is smaller than that of $\hat{\delta}_T$. As N increases and p approaches 0.5, the quantity on the right-hand side of (4.1) decreases and the MSE of $\hat{\delta}_T$ is likely to be smaller than that of $\hat{\delta}$, which can be checked in Figures 2 and 3.

5. The comparison of $\hat{\delta}_T$ and $\hat{\delta}$ in terms of the type I error rate

In this section, we compare the type I error rates of the two estimators ($\hat{\delta}_T$ and $\hat{\delta}$) in order to determine if the test treatment is effective in the target region ($\delta_T > 0$) when the treatment effects in both the target region and the non-target region are slightly different ($\delta_T \neq \delta_{NT}$).

5.1. Test based on $\hat{\delta}_T$

The hypothesis of interest for the target regulatory agency is expressed as (3.2). The test statistic based on $\hat{\delta}_T$ can be constructed as follows.

$$T_T = \frac{\hat{\delta}_T - \delta_T}{\sqrt{2S_{T,p}^2/n_T}},$$

where

$$S_{T,p}^2 = \frac{(n_T - 1)S_{T,X}^2 + (n_T - 1)S_{T,Y}^2}{2n_T - 2},$$

$$S_{T,X}^2 = \frac{1}{n_T - 1}\sum_{j=1}^{n_T}\left(X_{T,j} - \bar{X}_T\right)^2, \qquad S_{T,Y}^2 = \frac{1}{n_T - 1}\sum_{j=1}^{n_T}\left(Y_{T,j} - \bar{Y}_T\right)^2.$$

Although T_T follows a t distribution with $(2n_T - 2)$ degrees of freedom, we utilize the standard normal distribution as the null distribution of T_T, as $(2n_T - 2)$ is usually larger than 30. The rejection region of T_T is then the set which satisfies $T_T > 1.96$, and the type I error rate of T_T is approximately 2.5%.

5.2. Test based on $\hat{\delta}$

In this subsection, we test the treatment effect in the target region (δ_T) based on the overall treatment effect estimator ($\hat{\delta}$). The hypothesis of interest is still given by (3.2) and the test statistic can be devised as follows.

$$T_O = \frac{\hat{\delta} - \delta_T}{\sqrt{2S_p^2/N}} = \frac{p\hat{\delta}_T + (1-p)\hat{\delta}_{NT} - \delta_T}{\sqrt{2S_p^2/N}},$$

where

$$S_p^2 = \frac{(N-1)S_X^2 + (N-1)S_Y^2}{2N - 2},$$

$$S_X^2 = \frac{1}{N-1}\left[\sum_{j=1}^{n_T}\left(X_{T,j} - \bar{X}\right)^2 + \sum_{j=1}^{n_{NT}}\left(X_{NT,j} - \bar{X}\right)^2\right], \quad \bar{X} = \frac{1}{N}\left(\sum_{j=1}^{n_T}X_{T,j} + \sum_{j=1}^{n_{NT}}X_{NT,j}\right)$$

$$S_Y^2 = \frac{1}{N-1}\left[\sum_{j=1}^{n_T}\left(Y_{T,j} - \bar{Y}\right)^2 + \sum_{j=1}^{n_{NT}}\left(Y_{NT,j} - \bar{Y}\right)^2\right], \quad \bar{Y} = \frac{1}{N}\left(\sum_{j=1}^{n_T}Y_{T,j} + \sum_{j=1}^{n_{NT}}Y_{NT,j}\right).$$

In most MRCTs, we assume $\delta_T = \delta_{NT}$. Under such an assumption, the rejection region of T_O is the set which satisfies $T_O > 1.96$, and the type I error rate of T_O is approximately 2.5%.

Table 1: The type I error rate for $\hat{\delta}$ (%) ($\delta_T = 0$)

p	$\delta_{NT} = 0.05$			$\delta_{NT} = 0.01$	
	$N = 500$	$N = 1,000$		$N = 500$	$N = 1,000$
0.1	10.59	17.01		3.46	3.93
0.2	9.22	14.33		3.34	3.74
0.3	7.98	11.95		3.22	3.57
0.4	6.87	9.87		3.11	3.39
0.5	5.88	8.06		3.00	3.23
0.6	5.01	6.52		2.89	3.07
0.7	4.25	5.21		2.79	2.92
0.8	3.58	4.13		2.69	2.77
0.9	3.00	3.23		2.59	2.63
1.0	2.50	2.50		2.50	2.50

However, the main purpose of this paper is to investigate what would occur if $\delta_T \neq \delta_{NT}$ and the difference is small. When $\delta_T \neq \delta_{NT}$, the type I error rate of T_O for testing the hypothesis in (3.2) is as follows.

$$P\left(T_O > 1.96 | H_0 : \delta_T = 0, \delta_T \neq \delta_{NT}\right) = P\left(\frac{p\left(\hat{\delta}_T - 0\right) + (1-p)\left(\hat{\delta}_{NT} - \delta_{NT}\right) + (1-p)\delta_{NT}}{\sqrt{2S_p^2/N}} > 1.96\right)$$

$$\simeq 1 - \Phi\left(1.96 - \frac{(1-p)\delta_{NT}}{\sqrt{2\sigma^2/N}}\right).$$

Therefore, the type I error rate is greater than the nominal level when $\delta_T \neq \delta_{NT}$ and $\delta_{NT} > 0$. In other words, when the test treatment is effective only in the non-target region ($\delta_{NT} > 0$) and, not in the target region ($H_0 : \delta_T = 0$), if we test the treatment effect in the target region (δ_T) based on the overall treatment effect estimator ($\hat{\delta}$), the type I error rate is inflated. This is the price the target regulatory agency has to pay to borrow the information from the non-target region. This inflation of the type I error rate may be of concern to the target regulatory agency. On the other hand, if $\delta_T = 0$ and $\delta_{NT} < 0$, the type I error rate is deflated.

Table 1 shows the type I error rate of $\hat{\delta}$ when $\delta_T = 0$ for $\delta_{NT} = 0.05$ and 0.01. As the value of p increases from zero to one, the type I error rate approaches 2.5%. When $p = 1.0$, the target region is as identical to the overall population, and the type I error rate becomes 2.5%. Because the value of δ_{NT} is positive, the type I error rates are inflated for all values of p except $p = 1.0$. For example, when $p = 0.2$ and $\delta_{NT} = 0.05$ in Table 1, the type I error rates are 9.22% and 14.33% for $N = 500$ and $N = 1,000$ respectively. For even a small difference in the treatment effect, such as $\delta_T = 0$ and $\delta_{NT} = 0.05$, the inflation of the type I error rate is substantial. Note that $p = 0.2$ is one of the recommended sample sizes for the target region according to the Japanese guideline (MHLW, 2007).

6. Comparison of $\hat{\delta}_T$ and $\hat{\delta}$ in terms of power

In this section, we compare the powers of the two estimators ($\hat{\delta}_T$ and $\hat{\delta}$) in order to test the hypothesis in (3.2). Before we compare the powers of $\hat{\delta}_T$ and $\hat{\delta}$, it is necessary to consider the two factors below.

1. In order to investigate the power for testing the hypotheses in (3.2), we should assume that $\delta_T > 0$.

2. The main purpose of this paper is to investigate what would occur if $\delta_T \neq \delta_{NT}$ and the difference is small.
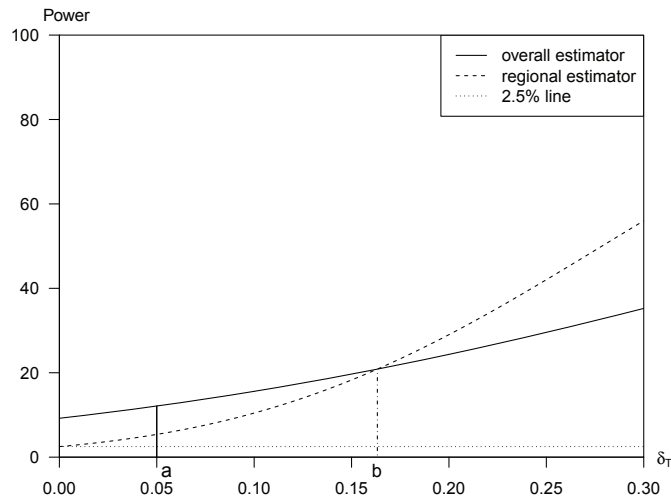
Figure 4: *The powers of $\hat{\delta}_T$ and $\hat{\delta}$ at $\delta_{NT} = 0.05$ and $p = 0.2$.*

Therefore, we compare the powers of $\hat{\delta}_T$ and $\hat{\delta}$ with two different methods.

6.1. Power as the value of δ_T increases

In this subsection, the powers of $\hat{\delta}_T$ and $\hat{\delta}$ are computed as the value of δ_T changes when the values of δ_{NT}, p, and N are fixed. For illustration purposes, the powers of $\delta_{NT} = 0.05$, $p = 0.2$, and $N = 500$ are presented in Figure 4. The powers of $\hat{\delta}_T$ and $\hat{\delta}$ increase as the value of δ_T increases. In Figure 4 $\delta_T = 0.05$ and $\delta_T = 0.163$ are denoted by a and b on the x-axis, respectively. Figure 4 shows that the powers of $\hat{\delta}$ are greater than those of $\hat{\delta}_T$ in the range of $0 < \delta_T < 0.163$, as the sample size of $\hat{\delta}$ is larger than that of $\hat{\delta}_T$ through the borrowing of information from the non-target region.

However, when $\delta_T > 0.163$, the power order is reversed, indicating that the powers of $\hat{\delta}_T$ are greater than those of $\hat{\delta}$ despite the fact that the sample size of $\hat{\delta}$ is larger than that of $\hat{\delta}_T$. This occurs because the value of δ_T ($\delta_T > 0.163$) is much greater than the value of δ_{NT} ($\delta_{NT} = 0.05$). Although $\hat{\delta}$ borrows information from the non-target region, this information does not contribute much to the increase in power, because $\delta_T > \delta_{NT}$.

6.2. Power as the value of N increases

In this subsection, the powers of $\hat{\delta}_T$ and $\hat{\delta}$ are computed as the sample size (N) changes when the values of δ_{NT}, p, and δ_T are fixed. In Figure 5 the powers are calculated in the four cases $(\delta_T, \delta_{NT}) = (0.05, 0.1)$, $(0.1, 0.1)$, $(0.2, 0.1)$, and $(0.3, 0.1)$ when $p = 0.2$. Given that $p = 0.2$, the sample size of the target region is smaller than that of the non-target region. Therefore, even when $(\delta_T, \delta_{NT}) = (0.1, 0.1)$, the powers of $\hat{\delta}$ are larger than those of $\hat{\delta}_T$. As the value of δ_T changes from 0.05 to 0.3, the difference in the powers of $\hat{\delta}_T$ and $\hat{\delta}$ is reduced.

From Figure 5, we find that the powers of $\hat{\delta}$ are greater than those of $\hat{\delta}_T$ unless δ_T is much larger than δ_{NT}. This is the main advantage of the overall treatment effect estimator ($\hat{\delta}$) compared to the regional treatment effect estimator ($\hat{\delta}_T$). This advantage is realized by sacrificing the type I error, as discussed in Section 4.
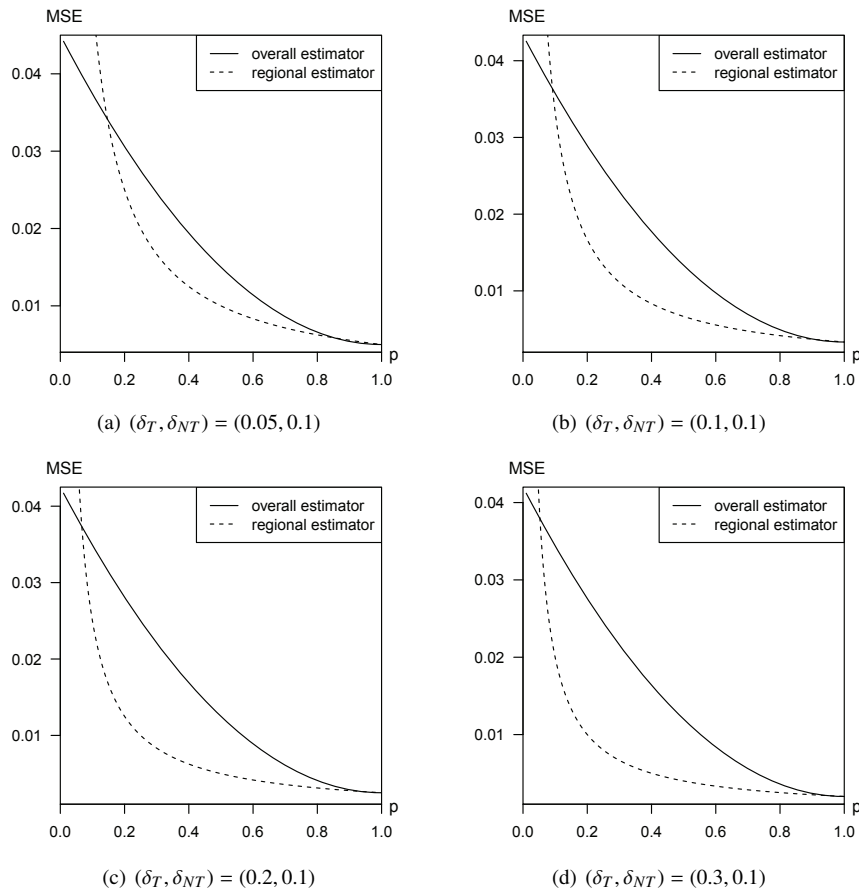
(a) $(\delta_T, \delta_{NT}) = (0.05, 0.1)$

(b) $(\delta_T, \delta_{NT}) = (0.1, 0.1)$

(c) $(\delta_T, \delta_{NT}) = (0.2, 0.1)$

(d) $(\delta_T, \delta_{NT}) = (0.3, 0.1)$

Figure 5: *The powers of $\hat{\delta}_T$ and $\hat{\delta}$ at $p = 0.2$.*

7. Discussion and concluding remarks

One of the problems associated with MRCTs is that there is no a global approval despite the fact that there is a pre-specified confirmatory analysis in the protocol. Therefore, each regional regulatory agency must make its own decisions based on its own analyses given an identical set of data from a MRCT. Under this situation, in this paper we compare two estimators ($\hat{\delta}$ and $\hat{\delta}_T$) which can be used to estimate the treatment effect in the target region (δ_T).

If the treatment effects in both the target region and the non-target region are completely identical ($\delta_T = \delta_{NT}$), the overall treatment effect estimator ($\hat{\delta}$) is more efficient than the regional treatment effect estimator ($\hat{\delta}_T$), which means that the power of $\hat{\delta}$ is greater than that of $\hat{\delta}_T$ while the type I error rate of $\hat{\delta}$ is held under the nominal level.

On the other hand, if $\delta_T \neq \delta_{NT}$ and the difference is small, a conventional method to estimate the treatment effect in the target region is to extrapolate the overall result to that region. In this paper, we investigate the statistical implications of such an extrapolation. Specifically, we show that, if the total sample size (N) is sufficiently larger than the sample size in the target region (n_T) and $|\delta_T - \delta_{NT}|$ is small, then the MSE of $\hat{\delta}$ is smaller than that of $\hat{\delta}_T$ despite the fact that $\hat{\delta}$ is not unbiased estimator of

δ_T. We also show that the powers of $\hat{\delta}$ are greater than those of $\hat{\delta}_T$ in exchange for inflating the type I error rates unless δ_T is much larger than δ_{NT}.

Another important issue is to distinguish "substantial" in the Section 2.2 from "small and clinically not meaningful" in the Section 2.3. This requires clinical judgement. For convenience of explanation, let us consider a hypothetical weight loss test treatment. Suppose that this weight loss test treatment reduces weight by 0.5 kg compared to placebo. Since 0.5 kg is not mathematically zero, it is true that the test treatment definitely reduces weight. However, the amount of weight lost is so small that no medical doctor will prescribe the test treatment. In other words, the test treatment has no clinical significance. Therefore, a new weight loss test treatment would have clinical significance only if weight reduction was sufficient. The question is how much weight should a new weight loss test treatment reduce in order to have clinical significance. This reference point is determined by clinicians, not statisticians. The problem is that the opinions of clinicians may differ.

Now, let us consider the problem of comparing δ_T and δ_{NT}. Whether the difference between δ_T and δ_{NT} is substantial is also a matter of clinical judgement, not statistical judgement. A problem is that opinions may differ between clinicians. There may be controversy in setting this value.

Anyhow, suppose that there is a positive value d such that the difference between δ_T and δ_{NT} is decided to be substantial if $\delta_T - \delta_{NT} > d$. Since δ_T and δ_{NT} are unknown parameters, there is no way to know for sure if $\delta_T - \delta_{NT} > d$ is satisfied. However, using the following statistical method may reduce the probability of making an erroneous decision. If the lower bound of 95% confidence interval of $\delta_T - \delta_{NT}$ is greater than d, we conclude that the difference between δ_T and δ_{NT} is substantial.

If the value of d is much greater than zero and the sample size is huge, then the 95% confidence interval of $\delta_T - \delta_{NT}$ may fall between $(0, d)$ interval, and we can conclude that the difference between δ_T and δ_{NT} is small and clinically not meaningful. However, if such conditions are not satisfied, it is usually very difficult to distinguish statistically between "no difference" in the Section 2.1 and "small and clinically not meaningful" in the Section 2.3."

In addition to the fixed-effect model, the random-effect model has been studied for MRCTs (Chen *et al.*, 2012b; Hung *et al.*, 2010). However, only the fixed-effect model is considered in this paper. Although the random-effect model has some advantages, it is not considered in this paper due to the following reasons.

- It is very difficult to accept that regions in MRCTs are randomly sampled from infinitely many regions. Regions in MRCTs are usually chosen, for instance, by considering disease prevalence and economic factors such as market sizes and the cost of clinical trials.

- It is the treatment effect in the target region (δ_T) that attracts the most interest from the target regulatory agency. However, under the random-effect model, the treatment effect in the target region is a random variable and, not a fixed parameter. Hence, it is impossible to estimate the treatment effect in the target region, because it is a random variable, though it is possible to predict it. Furthermore, it is impossible to conduct statistical hypothesis testing for the treatment effect in the target region (δ_T).

Acknowledgement

References

Chen J, Quan H, Binkowitz B, Quyang SP, Tanaka Y, Li G, Menjoge S, and Ibia E (2010). Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review, *Pharmaceutical Statistics*, **9**, 242–253.

Chen J, Quan H, Gallo P, *et al.* (2011). Consistency of treatment effect across regions in multi-regional clinical trials, part 1: design considerations, *Drug Information Journal*, **45**, 595–602.

Chen X, Lu N, Nair R, Xu Y, Kang C, Huang Q, Li N, and Chen H (2012a). Decision rules and associated sample size planning for regional approval utilizing multi-regional clinical trials, *Journal of Biopharmaceutical Statistics*, **22**, 1001–1018.

Chen CT, Hung HMJ, and Hsiao CF (2012b). Design and evaluation of multi-regional trials with heterogeneous treatment effect across regions, *Journal of Biopharmaceutical Statistics*, **22**, 1037–1050.

Hung HMJ, Wang SJ, and O'Neill RT (2010). Consideration of regional difference in design and analysis of multi-regional trials, *Pharmaceutical Statistics*, **9**, 173–178.

Ikeda K and Bretz F (2010). Sample size and proportion of Japanese patients in multi-regional trials, *Pharmaceutical Statistics*, **9**, 207–216.

International Conference on Harmonization (2006). Q&A for ICH E5 Guideline on Ethnic Factors in the Acceptability of Foreign Data.

International Conference on Harmonization (2017). General Principles for Planning and Design of Multi-Regional Clinical Trials.

Kawai N, Chung-Stein C, Komiyama O, and Li Y (2008). An approach to rationalize partitioning sample size into individual regions in a multi-regional trial, *Drug Information Journal*, **42**, 139–147.

Ko FS, Tsou HH, Liu JP, and Hsiao CF (2010). Sample size determination for a specific region in a multi-regional trial, *Journal of Biopharmaceutical Statistics*, **24**, 870–885.

Liu JP, Chow SC, and Hsiao CF (2013). *Design and Analysis of Bridging Studies*, CRC Press, Boca Raton, FL.

Ministry of Health, Labor, and Welfare of Japan (2007). Basic Principles on Global Clinical Trials, from: http://www.pmda.go.jp/english/service/pdf/notications/0928010-e.pdf

Quan H, Li M, Chen J, *et al.* (2010a). Assessment of consistency of treatment effects in multiregional clinical trials, *Drug Information Journal*, **44**, 617–632.

Quan H, Zhao PL, Zhang J, Roessner M, and Aizawa K (2010b). Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance, *Pharmaceutical Statistics*, **9**, 100–112.

Quan H, Li M, Shih WJ, Ouyang SP, Chen J, Zhang J, and Zhao P (2013). Empirical shrinkage estimator for consistency assessment of treatment effects in MRCT, *Statistics in Medicine*, **32**, 1691–1706.

Quan H, Mao X, Chen J, Shih WJ, Ouyang SP, Zhang J, Zhao PL, and Binkowitz B (2014). Multi-regional clinical trial design and consistency assessment of treatment effects, *Statistics in Medicine*, **33**, 2191–2205.

Tsong Y, Chang WJ, Dong X, and Tsou HH (2012). Assessment of regional treatment effect in a multi-regional clinical trial, *Journal of Biopharmaceutical Statistics*, **22**, 1019–1036.

Tsou HH, Chow SC, Lan KKG, *et al.* (2010). Proposals of statistical consideration to evaluation of results for a specific region in multi-regional trials - Asian perspective, *Pharmaceutical Statistics*, **9**, 201–206.

Tsou HH, Hung HMJ, Chen YM, Huang WS, Chang WJ, and Hsiao CF (2012). Establishing consistency across all regions in a multi-regional clinical trial, *Pharmaceutical Statistics*, **11**, 295–299.

Uesaka U (2009). Sample size allocation to regions in a multi-regional trial, *Journal of Biopharmaceutical Statistics*, **19**, 580–594.