Chapter 16
A Two-Tier Procedure for Designing and Analyzing Medical Device Trials Conducted in US and OUS Regions for Regulatory Decision Making

Nelson Lu, Yunling Xu and Gerry Gray

Abstract The number of clinical trials conducted simultaneously in the USA and outside of the US (OUS) for medical device development has been increasing over the last decade. However, the presence of inherent regional differences in treatment effects poses a great challenge to the US regulatory agency's decision making. In this chapter, we propose a two-tier procedure for analyzing data from such trials for the US regulatory agency's decision making, allowing treatment effects to vary from region to region. We differentiate direct evidence from supporting evidence while using both to exemplify the advantage of such trials for the US regulatory agency's decision making. The contribution of the supporting evidence can be adjusted according to the expectation of the magnitude of regional differences and the statutory requirements in the USA. Examples are presented to illustrate the design and analysis based on our proposed procedure. Using the proposed two-tier procedure with an upfront explicit decision tree can increase the predictability and transparency of the regulatory decision making.

16.1 Introduction

In the past decade, more and more medical device sponsors have begun conducting clinical trials simultaneously in the USA and outside of the US (OUS) to support regulatory approval of their products in the USA. Such trials are referred as multi-regional clinical trials (MRCTs) in this chapter. Lu et al. (2011) reported that from 2006 to 2010, about 21 % (17/81) of approved premarket applications (PMAs) for therapeutic devices at the Center for Device and Radiological Health (CDRH) are based on MRCTs conducted in the USA and OUS. Both sponsors and the Food and Drug Administration (FDA) are embracing such a concept, hoping to speed up medical device development, and thus to provide earlier availability of effective medical

Y. Xu (✉) · N. Lu · G. Gray
Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA
e-mail: Yun-Ling.Xu@fda.hhs.gov

Z. Chen et al. (eds.), *Applied Statistics in Biomedicine and Clinical Trials Design*,
ICSA Book Series in Statistics, DOI 10.1007/978-3-319-12694-4_16

devices to patients in the USA. Nonetheless, statistical issues for design, conduct, monitoring, and analysis of medical device MRCTs are very challenging, especially in a regulatory setting. Such challenges may not be the same as the ones encountered in drug MRCTs due to the fundamental differences in the characteristics of the products and the regulatory requirements among regions.

In Sect. 2 of this chapter, we present the issues associated with the current framework for analyzing MRCTs for regulatory decision making regarding the US medical device approval. In Sect. 3, a two-tier procedure is proposed for analyzing MRCTs for regulatory decision making with close alignment with the US statutory requirements. Examples for analyzing medical device trials are presented in Sect. 4, followed by design considerations in Sect. 5. The chapter concludes with discussion in Sect. 6.

16.2 Issues with Current Practice in Analyzing Medical Device MRCTs for Regulatory Decision Making in the USA

The statutory requirements for medical devices' premarket approval may vary significantly in different jurisdictions. For example, in the European Union (EU), the demonstration of the device effectiveness is not required for the CE (*Conformité Européenne* or *Communauté Européenne*) marking (CE Mark 2012). For the approval of a Class III medical device in the USA, a reasonable assurance of safety and effectiveness has to be demonstrated as indicated by Section 513(a)(1)(C) of the Federal Food, Drug, and Cosmetic Act (FD and C Act). As set forth in the US FD & C Act 513(a)(2), the safety and effectiveness of a medical device should be determined: (A) with respect to the persons for whose use the device is represented or intended and (B) with respect to the conditions of use prescribed, recommended, or suggested in the labeling of the device (Food, Drug and Cosmetic Act 2012). These statutory provisions specify that a finding of reasonable assurance of safety and effectiveness must be supported by data relevant to the target population, and evaluated in light of the device labeling (Guidance on the Collection of Race and Ethnicity Data in Clinical Trials 2012).

Following the US statutory requirement, if a study is intended to eventually support a premarket submission in the USA, selected study subjects should adequately reflect the target population for the device. This means, ideally the study should be conducted in the USA. Apparently, not all subjects in an MRCT are from USA. Regardless of where a study is conducted, it should be relevant to understanding the safety and effectiveness of the device when used in US subjects with regard to subject demographics, standard of care, practice of medicine, and any cultural differences in terms of expectations regarding medical care. "The Secretary shall accept data from clinical investigations conducted outside of the United States, including the European Union, if the applicant demonstrates that such data are adequate under applicable standards."(FADASIA 2014)

Currently, statistical inference on the global estimate of treatment effect based on pooled data is often used for regulatory decision making in approval of a medical

device in the USA based on an MRCT. Following International Conference on Harmonization (ICH) E5 Q&A #11 (2012), data from all regions are pooled together for analysis through prespecified hypothesis testing with a formal decision rule, and the treatment effect consistency across regions is assessed in a post hoc manner without a formal decision rule. There are notable difficulties associated with this current practice in analyzing MRCTs for regulatory decision making. To facilitate the discussion of these issues, let us set up a cell-mean model as follows:

For a randomized controlled superiority MRCT, let k index region: 1, 2, . . ., K; l index treatment (t) and control (c); ${n_k}^l$ be the sample size in region k for treatment l; and N_k be the size of the intended population in region k; μ_k^l be the cell mean for the population in region k with treatment l; δ_k ($= {\mu_k}^t - {\mu_k}^c$) be the treatment effect in region k. In a cell-mean model, the inference on the global mean of treatment effect in an MRCT is essentially to test the following hypothesis, where a larger μ indicates a better result:

$$H_0 : ({n_1}^t/n.^t){\mu_1}^t + \ldots + ({n_K}^t/n.^t){\mu_K}^t \leq ({n_1}^c/n.^c){\mu_1}^c + \ldots + ({n_K}^c/n.^c){\mu_K}^c$$

i.e., a test of treatment effect averaged across regions with a weight of ${n_k}^l/n.^l$; where $n.^l = n_1^l + \ldots + {n_K}^l$ attached to the region k. Please observe that:

1. If ${n_k}^l$ is not proportional to N_k within the MRCT and $\delta_1 = \ldots = \delta_K$ does not hold, the inference on the global estimate by the above test is based on a sample that does not match the population in any local region.
2. If ${n_k}^l$ is proportional to N_k within the MRCT and $\delta_1 = \ldots = \delta_K$ does not hold, the inference on the global estimate by the above test is for an intended population in the whole area covered by all the participated regions, which, however, does not match the population for any local region.
3. If $\delta_1 = \ldots = \delta_K$ holds, the inference on the global estimate by the above test is for an intended population, which matches the intended population in each of the local regions.

From the above observation, current practice in analyzing MRCTs for a local regulatory decision making is valid for that region only if the treatment effect is consistent across regions. There have been several papers discussing statistical methods for assessing treatment effect consistency across regions, for example, Chen et al. (2010), Hung et al. (2010), Quan et al. (2010), and Chen et al. (2012). Nonetheless, in a traditional hypothesis testing framework, it is inherently difficult to prove that $\delta_1 = \ldots = \delta_K$. The power for detecting treatment effect inconsistency among regions is generally fairly low when a study is only powered for testing the overall treatment effect.

A challenging regulatory issue is that, under the current framework, the regulatory decision becomes less predictable and less transparent when facing an observed state of heterogeneity in treatment effects. We believe that such an issue could be addressed with a prespecified decision tree, and our proposed statistical procedure in the next section should be able to serve this purpose.

16.3 A Two-Tier Procedure

For regulatory decisions regarding medical devices in the USA, the statutory requirement is that effectiveness be evaluated for the intended population identified in the labeling (Food, Drug and Cosmetic Act 2012). The intended population is usually characterized by its unique intrinsic and extrinsic factors, such as demographics, standard of care, practice of medicine, and any cultural differences in terms of expectations regarding medical care. As some of the intrinsic and extrinsic factors could be treatment effect modifiers, the effectiveness of a medical device should be evaluated as an estimate of efficacy for the intended population in the USA conditional on its unique intrinsic and extrinsic factors. In other words, potential regional difference must be taken into account when a regulatory decision in the USA on a medical device approval is made using MRCT data.

For medical devices, it is well known that physician's skill and accessibility of high-tech equipment contribute significantly to the effectiveness of a device in use, and this varies from region to region (Campbell 2008; Rothwell 2005). Tanaka (2010) discusses several US regulatory examples where regional treatment differences exist, and Tsou et al. (2010) discuss treatment effect differences from country to country in Asia. A similar pattern is observed in drug applications, as Hung et al. (2010) commented that "We have seen that many MRCTs suggest that there are regional differences in effect estimates."

To account for potential regional difference into an upfront decision tree, we here attempt to recast the issue of assessment on consistency of treatment effects across regions to an issue of information borrowing. The task is to incorporate effectiveness information from the OUS regions into the US regulatory decision making with acknowledgment that treatment effects may vary among regions. We propose a two-tier procedure for decision making in the US medical device approval based on MRCTs. The procedure is outlined in the following and displayed in Fig. 16.1. For convenience, region 1 is designated as the USA, the region of interest.

Step 1: Using data solely from region 1, test for H_{01}: $\delta_1 = 0$.

If the p value (p1) is less than a critical value c1, declare a tier 1 success in region 1; otherwise,

if p1 is less than a threshold value π ($\pi \geq$ c1), go to step 2;

otherwise, declare a failure in region 1.

Step 2: Using data from all regions, test for the effect of the variable treatment in the model, which has main effects treatment and regions, and the interaction term of treatment by region.

If the p value (p2) is less than a critical value c2, declare a tier 2 success in region 1; otherwise, declare a failure in region 1.

In the proposed two-tier procedure, direct evidence for the effectiveness of the product is evaluated in step 1 using data from region 1 only; and if warranted, supporting evidence is provided in step 2 using data from all regions. The null hypothesis listed in step 2 is that there is no treatment effect for the medical device

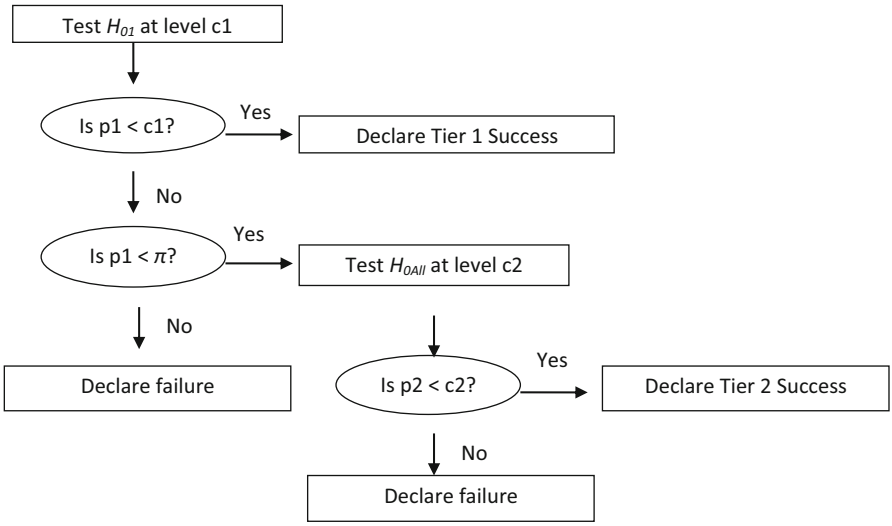

Fig. 16.1 Flow diagram of the two-tier procedure

based on a model that takes regions into consideration. For example, when testing a normally distributed endpoint, the null hypothesis is equivalent to H_{0all}: $\delta_1 + \delta_2 + \ldots \delta_K = 0$ when using type 3 sum of squares. In this case, the treatment effect is expressed as the average effectiveness across regions. Other statistical models may also be considered. Note that the distribution of the outcome measures can be of any type such as normal, binary, and censored failure time. A tier 1 success carries the most direct effectiveness evidence taking into account all the intrinsic and extrinsic factors associated with the intended population in the labeling; a tier 2 success is a synthesis of direct and supporting evidence from all regions by taking regional differences into account.

Note that the two-tier procedure (Fig. 16.1) is a decision procedure with an explicit decision tree, which can help the predictability of regulatory decision making in region 1. The false approval rate in region 1 is controlled at a level of c1 (for testing H_{01} in tier 1) plus $P_{H01,H0all}$ (c1 < p1 < π, p2 < c2) (for testing H_{0all} in tier 2). A desirable false approval rate in region 1 could be controlled by appropriately chosen c1, π, c2, and group sample size in all regions (n_k^l). In analyzing and designing an MRCT with the two-tier procedure, these choices are of paramount importance to interpretation of the trial result and should be based on the totality of considerations from both statistical and local regulatory perspectives. In the following sections, the choices of these parameters will be discussed.

The CDRH often has certain requirements for a minimal US sample size for some products to ensure the applicability of the study conclusion to the USA. In general, substantial expected regional differences would warrant a substantial proportion of the total sample size allocated to the local region. Once $n_k^l/n_\cdot^l$ is decided, one possible alpha allocation to the direct evidence is $(n_k^l/n_\cdot^l)\,\alpha$, where α is the probability of false

Table 16.1 Average infarct size for treatment and control across regions

	Treatment		Control	
Region	US	Europe	US	Europe
Average	30	33	40	31
Standard deviation	19	18	18	20
Sample size	77	132	24	48

approval in region 1. This reflects the extent of importance of direct evidence needed in the regulatory decision making. A small c1 implies using supporting evidence through the global test for H_{0all} unless the direct evidence is quite strong, whereas a larger c1 implies emphasizing direct evidence from the local region unless supporting evidence is necessary. Given the choice of c1, the critical value c2 could then be conservatively set equal to $\alpha - c1$. Alternatively, c2 can be obtained by simulation; the task is to find c2* such that the equation $(c1 + P_{H01,H0all}(c1 < p1 < \pi, p2 < c2^*)) = \alpha$ is satisfied. If the derived c2* is greater than α, c2 can be set at α and c1* can be derived by satisfying $(c1^* + P_{H01,H0all}(c1^* < p1 < \pi,\ p2 < \alpha)) = \alpha$. The threshold π specified in two-tier procedure is a design parameter, which determines when to use the supporting evidence. If π is set equal to c1, supporting evidence will never be used in the local regulatory decision making. If π is set equal to 1, supporting evidence will always be used in local regulatory decision making. Note that, when π is set equal to 0.5, supporting evidence can be used as long as the point estimate of treatment effect of the local region exhibits the desired direction. Depending on the expectation of the magnitude of regional difference and the willingness to use supporting evidence, π should be set between c1 and 0.5, say 0.15. This relatively small value for π means that supporting evidence will only be used if the result from tier 1 is "marginally" significant. This allows for the use of supporting evidence when warranted, while ensuring that a negative or poor outcome in a local region will not be overcome by results from other regions.

16.4 Examples for Analyzing Medical Device Trials

In this section, we illustrate how to analyze MRCTs data using the two-tier procedure with two hypothetical medical device premarket applications (by regulatory policy, we are not allowed to use real cases here). The first was an example of a cardiovascular interventional trial, and the primary endpoint was infarct size. The trial was a two-arm, randomized controlled study, and it was conducted in two regions: USA and Europe. Randomization was stratified by region. The descriptive result of the trial is shown in Table 16.1.

Suppose that the proposed two-tier procedure served as the decision rule in the USA with the rate of false approval (α) being set at 0.025. Based on the sample size within each region (Table 16.1), the critical value c1 is set at 0.009 ($= (n_i^l/n^l)\alpha$)

Table 16.2 Observed clinical success rate for treatment and control across regions

	Treatment		Control	
Region	USA	Europe	USA	Europe
Clinical success rate	54.4 % (35/65)	52.4 % (37/71)	35.7 % (12/34)	26.7 % (10/36)

according to the prespecified rule. The threshold value π is set at 0.15 as suggested above.

A two-sample t-test for the null hypothesis $\delta_{us} (= \mu^t_{us} \mu^c_{us}) \geq 0$ resulted in a p value (p1) of 0.0073, which is less than 0.009. Therefore, a tier 1 success is claimed; the direct evidence is strong enough for claiming a study success.

The second was an example of an ablation catheter to treat atrial fibrillation, and the primary endpoint was clinical success at 12 months. The trial was a two-arm, randomized controlled study with a treatment to control ratio of 2:1, and it was conducted in two regions: USA and Europe. Randomization was stratified by region. The descriptive result of the trial is shown in Table 16.2.

Suppose that the proposed two-tier procedure served as the decision rule in the USA with the rate of false approval (α) being set at 0.025. Based on the sample size within each region (Table 16.2), set critical value c1 = 0.012 ($= (n^l_i / n^l_\cdot)\alpha$) according to the prespecified rule. The threshold value π is set at 0.15.

A two-sample t-test for the null hypothesis $\delta_{us} (= p^t_{us} - p^c_{us}) \geq 0$ using the US data only resulted in a p value (p1) of 0.041. As p1 is greater than c1 (0.012) but less than π (0.15), H_{0all}: $\delta_{US} + \delta_{EU} = 0$ is tested using both the US and EU data. The resulting p value (p2) is 0.002, which is less than 0.013 (α-c1). Therefore, a tier 2 success was claimed. That is, the marginally significant direct evidence plus significant supporting evidence would lead to the US approval for the device.

16.5 Design Considerations: Sample Size Planning and Operating Characteristics

With the traditional two-sample test assuming constant treatment effect across regions, the design for an MRCT is relatively straightforward. Instead, using the proposed two-tier procedure as a tool, the design for an MRCT requires careful considerations and extensive simulations. In this section, we first discuss the paradigm of sample size planning. Then, we illustrate the process with a hypothetical example.

Fig. 16.2 is a diagrammatic display of the process for planning the sample size of an MRCT.

Step 1: Define regulatory decision context The regulatory decision context is device specific, mainly considering the intended population and its public health impact in the USA. From our review experience, for some devices, the clinical performance may be highly dependent on surgeon skills, health care system, medical

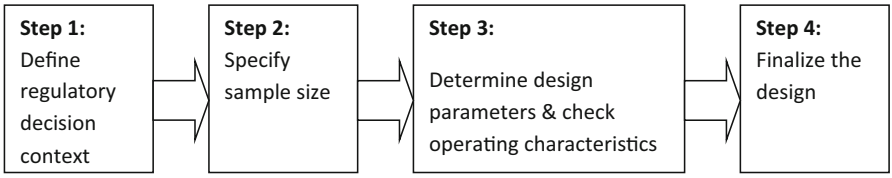

Fig. 16.2 Flow diagram of study design/sample size planning with the two-tier procedure

practice, and available ancillary surgical equipment in the country/region. In such cases, a larger sample size (or higher proportion) has often been called for with a consideration of the size of target population in the USA. Following our proposed paradigm, the direct evidence should be more valuable in the approval decision and thus the design parameter π should be set smaller. Considering that it is possible that the device works in other regions but has minimal effect in the USA, the false approval probability in the USA needs to be carefully considered.

Step 2: Specify sample size Within the defined regulatory decision context, the sample size can be specified with consideration of the sponsor's preference for the possible allocation of resources to the OUS regions.

Step 3: Determine design parameters and check operating characteristics Based on preliminary sample size allocation from step 2, the operating characteristics, such as the approval rates in the USA under different scenarios of true treatment effect in each region, δ_k, are examined via simulation. Meanwhile, the values of c1 and c2 (or c1* and c2*) will be determined per description in Sect. 3.

Step 4: Finalize the design There could be an iterative process between step 3 and step 2. When the statistical properties of the design, especially the false approval rate in the USA, are in alignment with the regulatory decision context and all stakeholders are in agreement, the sample size and all the design parameters are finalized.

In this hypothetical example, suppose that a two-arm, randomized controlled superiority MRCT is planned to be conducted in three regions (USA, region A, and region B) with a randomization ratio of 1:1 within each region. The clinical endpoint response follows a $N(\delta_k, 1)$ in the treatment arm and a $N(0, 1)$ in the control arm. A positive value of δ indicates a desirable outcome. Also, suppose that our proposed procedure is agreed upon between the CDRH and the sponsor.

A conventional way of designing such a trial serves as a good starting point. Assuming that the true treatment effect δ is 0.3 for all regions and that the data will be analyzed by pooling across regions, 174 subjects per arm are needed to have a power of 80 % with one-sided α of 0.025, using a two-sample t-test.

For illustration purpose, suppose that the CDRH calls for at least half (per regulatory decision context as discussed earlier in this section) of the sample size being from the USA. It is decided that the sample size is split roughly evenly in the other two OUS regions. Therefore, the sample size is allocated according to a 2:1:1 ratio,

or roughly 87, 44, and 43 subjects per arm in the USA, region A, and region B, respectively.

The two-tier procedure is implemented as the following. In tier 1, the two-sample t-test is performed using the data in the USA only. The analysis of variance (ANOVA) model is used in tier 2. The model includes main effects of region and treatment, along with the treatment $\times$ region interaction term. The p value of the Wald test for main effect term of treatment is obtained by PROC GENMOD of statistical analysis system (SAS), using type 3 sum of squares.

Please note that the operating characteristics are multifaceted due to the fact that the regional treatment effects are allowed to differ by region. For this allocated sample size, the design operating characteristics are examined based on nine different scenarios of δ_i's via simulation for illustration purpose here. (In practice, as many scenarios as desired should be evaluated). In the first five scenarios (A–E), the true treatment effect exists in the USA, while in the remaining scenarios (F–I), the true treatment effect is 0 in the USA. For each scenario, eight cases of values (c1, π, c2) are specified. These parameters are selected such that the false approval rate (in the USA) of scenario F is controlled at 0.025. The parameter π is set to range from 0.1 to 0.5. In cases b through g, the parameters are derived following our recommendation in Sect. 3. The value π is set at 0.1 in both cases a and b. Unlike case b where c1* and c2* are derived to control the false approval probability at 0.025, in case a, the c2 is conservatively set equal to $\alpha - c1$ (Table 16.3).

Several observations can be made by examining the simulation results. First, the approval rate based on t-test is fairly consistent at around 80 % when the overall average of the treatment effect is around 0.3, regardless of whether $\delta_1 = 0$, by comparing cases A, B, D, and E. This means that the t-test tends to inflate the false approval rate in the USA above the nominal alpha. Second, in scenarios G, H, and I, it is indeed shown that the false approval rate in the USA using t-test is higher than that using our proposed method. Third, when the device does work in the USA, the approval rate is generally getting larger with increasing π, except for scenario C. Meanwhile, with increasing π, the false approval rates in scenarios G, H, and I are increasing relatively rapidly than in other scenarios. This suggests that a smaller π may work better in controlling the false approval rate in the USA. Finally, the result in scenario D indicates that the proposed two-tier procedure has a higher approval rate than the t-test when the device is hardly effective in OUS regions.

Another set of simulation was done to investigate the impact of varying c1 (and thus c2) with a fixed value of π ($\pi = 0.15$, 0.2, and 0.3). The results, which are not presented here, indicate that the approval probabilities do not vary much.

Let us further examine some details from the extensive simulation for (c1, π, c2*) = (0.015, 0.15, 0.025). Note that the approval probability in scenario A based on the two-tier procedure is reduced from 80 to 71.6 % comparing to the conventional two-sample t-test, in which the treatment effect is assumed to be constant across regions. Taking into account of the potential differences in treatment effect across regions, the assumption of $\delta_1 = \ldots = \delta_K$ is relaxed in our proposed two-tier procedure. The reduction in the approval probability is mainly due to this relaxation. If it is desired to maintain the probability of approval at 80 % under the assumption of consistent treatment effect δ of 0.3, the sample size needs to be increased. Certainly,

Table 16.3 Simulation results on design operating characteristics based on t-test and two-tier procedure

Sample size (87:44:43)		Approval probability							
		t-test	Parameter set# of two-tier procedure						
Scenario	(δ_{US},δ_A,δ_B)		a	b	c	d	e	f	g
A	(0.3, 0.3, 0.3)	0.801	0.626	0.678	0.716	0.738	0.750	0.748	0.744
B	(0.3, 0.4, 0.2)	0.795	0.617	0.663	0.698	0.716	0.730	0.728	0.724
C	(0.3, 0.0, 0.0)	0.278	0.414	0.477	0.455	0.433	0.427	0.423	0.421
D	(0.6, 0.0, 0.0)	0.785	0.956	0.969	0.964	0.958	0.957	0.957	0.957
E	(0.3, 0.6, 0.0)	0.800	0.555	0.597	0.613	0.618	0.634	0.639	0.640
F	(0.0, 0.0, 0.0)	0.025	0.017	0.025	0.025	0.025	0.025	0.025	0.025
G	(0.0, 0.3, 0.3)	0.284	0.070	0.082	0.116	0.145	0.185	0.219	0.244
H	(0.0, 0.3, 0.6)	0.535	0.093	0.097	0.143	0.187	0.266	0.337	0.408
I	(0.0, 0.0, 0.6)	0.270	0.068	0.080	0.112	0.143	0.186	0.215	0.270

#:
a: $c1 = 0.025/2$, $\pi = 0.1$, $c2 = 0.025/2$
b: $c1^* = 0.018$,$\pi = 0.1$, $c2^* = 0.025$
c: $c1^* = 0.015$,$\pi = 0.15$, $c2^* = 0.025$
d: $c1 = 0.025/2$,$\pi = 0.2$, $c2^* = 0.025$
e: $c1 = 0.025/2$, $\pi = 0.3$, $c2^* = 0.02$
f: $c1 = 0.025/2$, $\pi = 0.4$, $c2^* = 0.017$
g: $c1 = 0.025/2$, $\pi = 0.5$, $c2^* = 0.015$

there are numerous ways to allocate the extra needed subjects, based on the requirement of the regulatory agency and the resources of the sponsor. Suppose that all extra subjects are determined to be assigned to the USA. Through a trial-and-error process of simulations, it can be found that a total of 118 subjects per arm are required in the USA to achieve a probability of approval of 80 %, when the design parameters ($c1^*$, π, $c2^*$) = (0.016, 0.15, 0.025).

In summary, evaluation of operating characteristics for a design with the two-tier procedure is inherently multifaceted as there are many ways to construct treatment effects varying across regions. A thorough exploring over many scenarios is of paramount importance to help understand the impact of anticipated and unexpected regional differences on the approval rate in the USA and to reach an agreement among stakeholders.

16.6 Discussion

Our proposed framework is devised to fit the situations that are common or relatively unique in the medical device trials. First, regulatory requirements for premarket approval may be different across regions, as discussed in Sect. 2. Consequently, the decision rule or the success criteria of a trial may be different across regions. Second,

the number of regions in many medical device MRCTs is relatively small. This may be due to the overall smaller sample size resulting from generally larger effect size of medical devices (as compared to drugs). In some cases, the number of regions is limited due to the accessibility of high-tech equipment and the requirement of innovative or delicate surgical techniques. Third, the consistency of the treatment effect may be suspicious even in the design stage.

While the proposed two-tier procedure provides an explicit decision tree upfront, it requires increased rigor to demonstrate effectiveness in the local region of interest, which can lead to a greater sample size. As the direct and supporting evidence are defined in terms of p values from statistical tests, the proposed procedure is perhaps more meaningful and works better when the sample size in the local region is relatively large. Motivated by our regulatory review experience, in this chapter, we attempt to develop a procedure for use in the USA by closely following the US medical device law and we have noticed that a large proportion of the sample size are from the USA in many submissions to the CDRH. Note that this two-tier procedure does not need to be adopted in every region even within the same MRCT as the medical device laws vary significantly from region to region. Alternatively, the proposed two-tier procedure can also work with relatively small sample size in a local region by setting π close to 1. Considering judiciary independence in medical device approvals across regions, each region could adopt its own statistical analysis plan.

In a regulatory setting, it is necessary to predefine the regions in an MRCT and ideally to have randomization stratified by region to facilitate the all-region analysis. The geographic area under the US FDA jurisdiction would form the main region for effectiveness evaluation; OUS regions could be predefined by various criteria. One is to be formed according to judicial areas. Another is to be formed across judicial boundaries according to similarity in intrinsic and extrinsic factors, such as medical practice and healthcare policy in particular, as discussed by Binkowitz (2010).

An important design feature with the two-tier procedure is the adjustability of acceptable levels of direct versus supporting evidence to meet regulatory expectations. When less regional treatment effect difference is expected, a regulatory decision could be based more on significant supporting evidence through setting the design parameter π closer to 0.5 from below; when substantial regional treatment effect difference is expected, a regulatory decision should be based less on significant supporting evidence through setting the design parameter π closer to alpha from above. In an MRCT with the two-tier procedure, the false approval rate for a region (say region A, $\delta_A = 0$) is evaluated upfront at design stage under many scenarios (δother-region be any plausible values) to understand the impact of plausible regional difference on regulatory decision making and subsequently help all the stakeholders reach an agreement on a study design.

In summary, our proposed two-tier procedure represents a new paradigm in which an explicit decision tree is generated upfront to increase the transparency and predictability for regulatory decision making in contrast to the current paradigm in which there is no explicit decision tree for regulatory decision making when the consistency of regional treatment effects is in doubt. We feel that it is better

aligned with the statutory requirements for medical device approval in the USA to have a regulatory decision making from analyses based on a careful evaluation to account for undesirable regional differences in treatment effect at the design stage.

Disclaimer No official support or endorsement by the Food and Drug Administration of this article is intended or should be inferred.

References

Binkowitz B (2010) Highlights from the PhRMA MRCT Key Issue Team & DIA MRCT Workshop. Presented at: the 4th Seattle Symposium in Biostatistics: Clinical Trials. Seattle, WA, 20–23 November

Campbell G (2008) Statistics in the world of medical devices: the contrast with pharmaceuticals. J Biopharm Stat 18:4–19

Chen J, Quan H, Binkowitz B et al (2010) Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. Pharm Stat 9:242–253

Chen J, Quan H, Gallo P et al (2012) An adaptive strategy for assessing regional consistency in multiregional clinical trials. Clin Trials 9:330–339

CE Mark (2012) http://eur-lex.europa.eu/LexUriServ/site/en/consleg/1993/L/01993L0042-20031120-en.pdf. Accessed 22 Oct 2012

FADASIA (2014) http://www.gpo.gov/fdsys/pkg/BILLS-112s3187enr/pdf/BILLS-112s3187enr.pdf. Accessed 3 Feb 2014

Food, Drug and Cosmetic Act (FD & C Act) (2012) www.fda.gov/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/FDCActChapterVDrugsandDevices/ucm110188.htm. Accessed 25 May 2012

Guidance on the Collection of Race and Ethnicity Data in Clinical Trials (2012) http://www.fda.gov/RegulatoryInformation/Guidances/ucm126340.htm. Accessed 22 Oct 2012

Hung HMJ, Wang S-J, O'Neill RT (2010) Consideration of regional difference in design and analysis of multi-regional trials. Pharm Stat 9:173–178

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 (2012) www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E5_R1/Q_As/E5_Q_As__R5_.pdf. Accessed 25 May 2012

Lu N, Nair R, Xu Y (2011)Decision rules and associated sample size planning for regional approval utilizing multi-regional clinical trials. Presented at: the 4th Annual FDA/MTLI Medical Device and IVD Statistical Issues Workshop, National Harbor, MD, 13-14 April, 2011

Quan H, Chen J, Gallo P et al (2010) Assessment of consistency of treatment effects in multiregional clinical trials. Drug Inf J 44:617–632

Rothwell P (2005) External validity of randomized controlled trials: to whom do the results of this trial apply?. Lancet 365:82–93

Tanaka Y (2010) Statistical considerations in multi-regional clinical trials. Presented at: the Biopharmaceutical Applied Statistics Symposium XVII, Hilton Head, SC, 5–9 November, 2010

Tsou H-H, Chow S-C, Lan KKG et al (2010) Proposals of statistical consideration to evaluation of results for a specific region in multi-regional trials—Asian perspective. Pharm Stat 9:201–206