Published in final edited form as:

J Biopharm Stat. 2017 ; 27(6): 933–944. doi:10.1080/10543406.2017.1293075.

Statistical design of noninferiority multiple region clinical trials to assess global and consistent treatment effects

Guoqing Diao^a, Donglin Zeng^b, Joseph G. Ibrahim^b, Alan Rong^c, Oliver Lee^c, Kathy Zhang^c, and Qingxia Chen^d

^aDepartment of Statistics, George Mason University, Fairfax, Virginia, USA

^bDepartment of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

^cAmgen Inc., Thousand Oaks, California, USA

^dDepartment of Biostatistics, Vanderbilt University, Nashville, Tennessee, USA

Abstract

Noninferiority multiregional clinical trials (MRCTs) have recently received increasing attention in drug development. While a major goal in an MRCT is to estimate the global treatment effect, it is also important to assess the consistency of treatment effects across multiple regions. In this paper, we propose an intuitive definition of consistency of noninferior treatment effects across regions under the random-effects modeling framework. Specifically, we quantify the consistency of treatment effects by the percentage of regions that meet a predefined treatment margin. This new approach enables us to achieve both goals in one modeling framework. We propose to use a signed likelihood ratio test for testing the global treatment effect and the consistency of noninferior treatment effects. In addition, we provide guidelines for the allocation rule to achieve optimal power for testing consistency among multiple regions. Extensive simulation studies are conducted to examine the performance of the proposed methodology. An application to a real data example is provided.

Keywords

Consistency of treatment effects; global treatment effect; multiregional clinical trial; noninferiority clinical trial; random effects model; signed likelihood ratio test

1. Introduction

Drug development in recent years has seen the increased popularity of multiregional clinical trials (MRCTs). By including relevant regions in one protocol, MRCTs, when well conducted, can simultaneously achieve two important goals, per the International Conference on Harmonization (ICH) E5 and ICH E17 (ICHE17) guidelines (ICH1998). First, an MRCT can provide a global estimate of treatment effect by combining data from

different regions. Secondly, it can provide the needed regional information to bridge the global conclusion to each region or between different regions (Hung et al., 2010). However, the key to achieving these goals rests mainly on the consistency of the treatment effects across different regions. If the treatment effects are similar across regions, the global treatment effect estimate should provide a more precise estimate for the treatment effect while regional estimates should provide adequate evidence for internal bridging. Otherwise, reasons for regional differences need to be explored and additional bridging studies may be necessary.

For each MRCT, evaluation of treatment consistency usually covers both the global consistency and the regional-specific consistency. Although regional-specific consistency (one region vs. the rest, or one region vs. another) is relevant for regional approval, global consistency (whether treatment effects are similar across all regions) is more essential to warrant the validity of the primary study conclusion. Several regional agencies, including the Committee for Medicinal Products for Human Use Committee for Medicinal Products for Human Use (CHMP) (2009) and Ministry of Health, Labour and Welfare of Japan (2007), have issued guidance on the assessment of treatment consistency in MRCTs. More recently, Quan et al. (2010), Chen et al. (2010), and Shih and Quan (2013) described a variety of methods for evaluating the global and/or regional treatment consistency. A summary of the regulatory requirements and detailed discussion of these methods are provided in the book *Design and Analysis of Bridging Studies* (Liu et al., 2013).

Most MRCTs assume treatment consistency when designing the trial. But in reality, regional treatment effects differ due to potential intrinsic and extrinsic factors as classified by ICH E5 Guideline such as ethnics, culture, diet, genetics, and environment. See Chapter 1 of *Design and Analysis of Bridging Studies* (Liu et al., 2013) for a detailed discussion. A fixed effects model as well as a random effects model can provide estimates on both overall and regional treatment effects. The question as to which model is more appropriate often remains open. There is however a growing recognition that a random effects model may provide more robust assessments in MRCTs due to its capacity to incorporate the additional between-region variability in the estimation procedure. Furthermore, a fixed-effects model may involve many parameters when there are many regions. See Hung et al. (2010) for a discussion on how different models can impact study design on power and sample size, and Quan et al. (2013) on how to use the random effects model to estimate the overall and regional treatment effect. As active treatments are approved and available to patients, placebo controlled studies for the development of additional treatments may no longer be ethical.

As pointed out in Section 3.5.2 by Shih and Quan (2013), most of the methods discussed so far have focused consistency on superiority trials hence may not be directly applicable to noninferiority trials. A noninferiority study compares an investigational drug with an active control with the objective of demonstrating that the difference between the two is within an acceptable limit. The experimental drug may have nonefficacious benefits such as reduced toxicities, lower cost, or route of administration. Final guidance on the conduct of noninferiority clinical trials was provided by FDA (2016). Methods to define the noninferiority margin and critical assumptions of noninferiority studies are all described in the FDA

document. There is limited work on assessing consistency in a noninferiority MRCT. Proposed methods for non-inferiority trials include shifted quantitative and qualitative treatment by region interaction tests (Wiens and Heyse, 2003; Hsieh and Liu, 2013) and a simultaneous test by Quan and Shih (2002) that jointly tests all regions against region-specific noninferiority margins. Essentially all these methods are based on fixed effect models.

A motivating example was a noninferiority MRCT on patients with colorectal cancer (CRC), which is the fourth-leading cause of cancer-related deaths worldwide. It has been known that the anti-epidermal growth factor receptor (EGFR) monoclonal antibodies cetuximab (a chimeric IgG1 antibody) and panitumumab (a fully human IgG2 antibody) provide clinical benefit in patients with chemotherapy-refractory metastatic colorectal cancer (mCRC) in the monotherapy setting (Jonker et al., 2007; Van Cutsem et al., 2007). The standard of care has evolved since the approval of these agents, and consequently, the re-evaluation of safety and efficacy in this setting is important and timely. The randomized, double-blind, phase 3 ASPECCT study (A Study of Panitumumab Efficacy and Safety Compared to Cetuximab) is the first head-to-head comparison of overall survival (OS) benefit of panitumumab and cetuximab in chemotherapy-refractory wild-type Kirsten rat sarcoma viral oncogene homolog exon 2 mCRC (Price et al., 2014). ASPECCT, which enrolled patients worldwide, is the largest prospective direct comparison of anti-EGFR therapies in mCRC, and is among the largest head-to-head comparisons of anticancer biologic agents in mCRC. The study was conducted in 27 countries in North and South America, Europe, Asia, Africa, and Australia. Four countries with sample sizes ≤ 5 were excluded from the analysis. Crossover between panitumumab/cetuximab was not permitted during the study treatment period. Patients were followed up for OS for 24 months after randomization of the last patient. Given the similar mechanisms of action, ASPECCT is a noninferiority study designed to demonstrate that panitumumab retains $\geq 50\%$ of the OS treatment effect of cetuximab versus best supportive care (BSC). The predicted effect of cetuximab versus BSC on OS was based on the CO.17 study (Karapetis et al., 2008) (hazard ratio [HR]= 0.55; 95% CI, 0.41–0.74). A thousand patients were enrolled to achieve 90% power with a one-sided $\alpha = 0.025$ for the OS inferiority null hypothesis.

In this paper, we propose a statistical design for an MRCT in order to evaluate the noninferiority of an experimental drug in multiple regions and to assess consistency of the treatment effects across regions. We provide a new definition of consistency of noninferior treatment effects across multiple regions under the random effects modeling framework. We then propose to use the signed likelihood ratio test (LRT) for testing consistency and derive an optimal allocation rule to maximize the power. Extensive simulation studies are carried out to examine the performance of the proposed methodology. An application to a real data example is provided. The rest of this paper is organized as follows. In Section 2, we describe the proposed methodology to define and test consistency of treatment effects in a noninferiority MRCT. Extensive simulations studies are presented in Section 3 to examine the finite sample properties of the proposed methodology. In Section 4, we illustrate the new methods through an application to the motivating ASPECCT study. We conclude with a brief discussion in Section 5.

2. Methods

2.1. Definition of consistency

In this section, we provide an intuitive and new definition of consistency of noninferior treatment effects across regions. We focus on global consistency because it is most important. Noninferiority MRCTs, with their usual large study size, often enroll patients in large number of regions. Unless specifically committed, sample sizes in most regions (if not all) in noninferiority MRCTs do not warrant an independent region-specific test of noninferiority with reasonable power. Therefore, almost all regulatory agencies will rely heavily on strength of the global consistency of the data to extrapolate the overall study conclusion to their specific regions for drug approval.

Consider L regions and let β_l denote the log-hazard ratio for the l th region, $l = 1, \dots, L$. Since β_l varies from region to region, a reasonable model to quantify this variability is to assume a random effects model:

$$\beta_l = \beta + u_l, \quad l = 1, \dots, L, \quad (1)$$

where the u_l 's are i.i.d $N(0, \sigma^2)$ random variables and σ^2 characterizes the variability across the regions. This random-effects model framework allows one to model heterogeneity among the patient population across different regions. A value of 0 for σ^2 implies that there is no treatment difference across different regions. This kind of mixed effects model has been extensively used in different settings such as in meta-analysis.

In model (1), β represents the overall treatment effect across regions. Therefore, a noninferiority MRCT involves testing $H_{0g}: \beta - \beta_0 > M$, where β_0 is the effect of an active control and M is the specified margin. In practice, we usually set $\beta_0 = 0$. Rejecting H_{0g} will provide evidence to declare that the log-hazard of the experimental drug is at most M larger than the log-hazard of the active control. In an MRCT, in addition to testing the treatment effect, another important objective is to assess the global consistency across regions. Here we propose an intuitive definition for global consistency using a probabilistic approach to accommodate the usual large number of regions involved in noninferiority MRCTs.

Assuming all regional treatment effects come from the same underlying distribution, we define consistency based on the percent of regions among all regions that indeed reach noninferiority. Consistency is achieved when the percentage of those regions surpasses a prespecified threshold. Notice this definition is different from the concept of treatment similarity proposed by Liu et al. (2009) for superiority trials although both involve noninferiority hypothesis testing. Under this definition, consistency is achieved as long as the treatment is noninferior to the active control in the majority of the regions even if the regional effects are very different.

For the above consistency definition, one quantity for assessing the majority of noninferior treatment effects is $P(\beta_l - \beta_0 \leq M)$ where β_l is given in model (1). Therefore, if this probability is larger than a prespecified proportion p (say, 80%), then this means that at least $p \times 100\%$ of the regions would have noninferior treatment effects. Since $P(\beta_l - \beta_0 \leq M) = \Phi$

$\{(\beta_0 + M - \beta)/\sigma\}$ where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, consistency thus holds if $\beta - \beta_0 \leq M - \sigma \Phi^{-1}(p)$, where $\Phi^{-1}(p)$ is the 100 p th percentile of the standard normal distribution. In other words, the null hypothesis we wish to reject is:

$$H_{0c}: \beta - \beta_0 > M - \sigma \Phi^{-1}(p).$$

The above definition of consistency enables us to test both the global treatment effect and the consistency of treatment effects across multiple regions with one test and one common noninferiority margin. Note that the null hypothesis for testing the global treatment effect is:

$$H_{0g}: \beta - \beta_0 > M$$

and $\Phi^{-1}(0.5) = 0$. Therefore, to test the global treatment effect is equivalent to testing that 50% of the regions have noninferior treatment effects by setting $p = 0.50$. If the margin M is deemed to be too rigorous, we can certainly use a separate test with a different (and perhaps more relaxed) margin to test the hypothesis for global consistency.

2.2. Signed (LRT) approach

In this section, we describe a signed LRT (Barndorff-Nielsen, 1986) for testing the consistency of noninferior treatment effects defined in the previous section 2.1. Without loss of generality, we fit a survival model for each region separately and obtain the parameter estimates $\hat{\beta}_l$, $l = 1, \dots, L$. The estimated variance of $\hat{\beta}_l$ based on the data from the l th region is $\hat{\sigma}_l^2$, $l = 1, \dots, L$. We then fit the following linear mixed model:

$$\hat{\beta}_l = \beta + u_l + \varepsilon_l, \quad l = 1, \dots, L, \quad (2)$$

where $u_l \sim N(0, \sigma^2)$ and $\varepsilon_l \sim N(0, \hat{\sigma}_l^2)$. We assume $\hat{\sigma}_l^2 = \hat{v}_l/n_l$ is known, where $\hat{v}_l$ is related to disease prevalence of each specific region (Kalbfleisch and Prentice, 2011, Chapter 5.7) and is often available to investigators. We can maximize the following log-likelihood to obtain the maximum likelihood estimators (MLEs) of β and σ , denoted by $\hat{\beta}$ and $\hat{\sigma}$:

$$\log L(\beta, \sigma) = \sum_{l=1}^L -\frac{1}{2} \log(\sigma^2 + \hat{\sigma}_l^2) - \sum_{l=1}^L \frac{(\hat{\beta}_l - \beta)^2}{2(\sigma^2 + \hat{\sigma}_l^2)} + C,$$

where C is a constant.

Recall that the null hypothesis H_{0c} (i.e., $p \times 100\%$ of the regions have noninferior treatment effects) for testing consistency is given by:

$$H_{0c}: P(\beta_l - \beta_0 \leq M) \leq p,$$

where $\beta_l = \beta + u_l$. We can rewrite the above null hypothesis as:

$$H_{0c}: \gamma = \beta_0 + M,$$

where $\gamma = \beta + \sigma \Phi^{-1}(p)$. Letting $\gamma_0 = \beta_0 + M$, the alternative hypothesis H_{1b} is given by:

$$H_{1b}: \gamma < \gamma_0$$

We propose to use a profile likelihood approach. The log-likelihood of the unknown parameters $\theta = (\gamma, \sigma^2)$ is given by:

$$\log L(\gamma, \sigma) = \sum_{l=1}^L -\frac{1}{2} \log(\sigma^2 + \hat{\sigma}_l^2) - \sum_{l=1}^L \frac{(\hat{\beta}_l - \gamma + \sigma \Phi^{-1}(p))^2}{2(\sigma^2 + \hat{\sigma}_l^2)}.$$

We denote the MLE of θ under the null hypothesis H_{0b} and the alternative hypothesis H_{1b} by $\tilde{\theta} = (\gamma_0, \tilde{\sigma}^2)$ and $\hat{\theta} = (\hat{\gamma}, \hat{\sigma}^2)$, respectively. The LRT statistic for testing H_{0b} is given by:

$$W = -2 \left\{ \log L(\tilde{\theta}) - \log L(\hat{\theta}) \right\}.$$

Under the null hypothesis, W asymptotically follows a χ_1^2 distribution. Since we have a one-sided alternative hypothesis, we consider the signed likelihood ratio test statistic (Barndorff-Nielsen, 1986):

$$R_{MLE} = \text{sign}(\hat{\gamma} - \gamma_0) \sqrt{W},$$

where $\text{sign}(x) = I(x > 0) - I(x < 0)$ and $I(\cdot)$ is the indicator function. Under H_{0b} , R_{MLE} asymptotically follows a standard normal distribution and we reject the null hypothesis for small values of R_{MLE} .

It is known that the MLE of σ^2 is negatively biased especially when the sample size is small. Alternatively, we can obtain the restricted maximum likelihood (REML) estimator of θ by maximizing the restricted log-likelihood:

$$\log L_{REML}(\gamma, \sigma^2) = \log L(\gamma, \sigma^2) - \frac{1}{2} \log \sum_{l=1}^L \frac{1}{\sigma^2 + \hat{\sigma}_l^2}.$$

Similarly, we can calculate LRT statistic using the REML estimators, leading to:

$$W_{REML} = -2 \left\{ \log L_{REML}(\tilde{\theta}_{REML}) - \log L_{REML}(\hat{\theta}_{REML}) \right\},$$

where $\hat{\theta}_{REML}$ and $\hat{\gamma}_{REML}$ are the REML estimators under the null and alternative hypotheses, respectively. The corresponding signed LRT statistic is defined as:

$$R_{REML} = \text{sign}(\hat{\gamma}_{REML} - \gamma_0) \sqrt{W_{REML}},$$

where $\hat{\gamma}_{REML}$ is the REML estimator of γ under H_{1b} .

Both signed LRT statistics R_{MLE} and R_{REML} have the same asymptotic standard normal distribution. However, the test statistic R_{REML} based on the REML estimators is expected to perform better than R_{MLE} with a small sample size, which is typical in multiple region clinical trials. We use the routine `optim()` in R (R Core Team, 2013) to maximize the log-likelihood and the restricted log-likelihood. The proposed signed LRT tests have been implemented in an R package, which is freely available upon request.

2.3. Optimal allocation

In this section, we provide a procedure to allocate patients to achieve optimal power for testing the consistency among multiple regions in a noninferiority clinical trial. Let π_l be the proportion of subjects in the l th region, $l = 1, \dots, L$. We propose to minimize the limiting

variance of $\hat{\gamma}$ for given p , v_l , $l = 1, \dots, L$, and the total sample size $N \equiv \sum_{l=1}^L n_l$. By inverting the Fisher information matrix, we can show that:

$$\text{Var}(\hat{\gamma}) \approx \left(\sum_{l=1}^L \frac{1}{\sigma^2 + v_l/n_l} \right)^{-1} + \left\{ \Phi^{-1}(p) \right\}^2 \left\{ \sum_{l=1}^L \frac{2\sigma^2}{(\sigma^2 + v_l/n_l)^2} \right\}^{-1},$$

where $n_l = N\pi_l$, $l = 1, \dots, L$. We minimize the above variance under the constraint

$\sum_{l=1}^L \pi_l = 1$ by using the quasi-Newton algorithm. It is worth noting that in order to obtain the above allocation rule, one needs to specify the values of σ^2 , p , v_l , $l = 1, \dots, L$ and the total sample size N .

2.4. A special case

In this section, we consider a special case under which the within region variances (v_l 's) are assumed to be the same, i.e., $v_1 = \dots = v_L = v$. In this case, it can be shown that the equal allocation rule is optimal. In fact, we can show that the minimizer of:

$$\left(\sum_{l=1}^L \frac{1}{\sigma^2 + v_l/n_l} \right)^{-1} + \left\{ \Phi^{-1}(p) \right\}^2 \left\{ \sum_{l=1}^L \frac{2\sigma^2}{(\sigma^2 + v_l/n_l)^2} \right\}^{-1}$$

is $\pi_1 = \dots = \pi_L = 1/L$, $l = 1, \dots, L$.

Define $\sigma_e^2 = Lv/n$. Under regularity conditions and assuming $\sigma > 0$, we can show that:

$$\frac{\sqrt{L}(\hat{\gamma} - \gamma)}{\sqrt{(\sigma^2 + \sigma_e^2) + \{\Phi^{-1}(p)\}^2 \frac{(\sigma^2 + \sigma_e^2)^2}{2\sigma^2}}} \rightarrow N(0, 1)$$

in distribution as $L \rightarrow \infty$. Therefore, when $\gamma - \gamma_0 = L^{-1/2}\delta$, the asymptotic local power of the signed LRT test at significance level α is $\Phi(-Z_\alpha - \delta^*)$, where:

$$\delta^* = \frac{\delta}{\sqrt{(\sigma^2 + \sigma_e^2) + \{\Phi^{-1}(p)\}^2 \frac{(\sigma^2 + \sigma_e^2)^2}{2\sigma^2}}}.$$

In our experience, the above large sample approximation works reasonably well when L is larger than 20. We note that given L and σ^2 , the above asymptotic local power reaches its upper bound when $n \rightarrow \infty$, under which:

$$\delta^* = \frac{\delta}{\sigma \sqrt{1 + \{\Phi^{-1}(p)\}^2 / 2}}.$$

3. Simulation studies

We conduct extensive simulation studies to examine the small sample performance of the proposed signed LRTs. We generate data from model (2) using settings similar to the ASPECCT study. Specifically, we set the values of L , v_k and π_k , $k = 1, \dots, L$ from the ASPECCT study as presented in Table 1, where each of the 23 countries with enrollment of 10 subjects in this study defines one specific region. We also set $\beta_0 = 0$ and $M = 0.15$. We vary the total sample size N and the variance parameter σ^2 such that $N = j \times 575$, $j = 1, \dots, 4$, and $\sigma = k \times 0.1345$, $k = 1, \dots, 4$. Here the case of $N = 1150$ and $\sigma = 0.269$ corresponds to the setting obtained from the ASPECCT study. Note that while the null value of γ is fixed, different values of σ lead to different null values of β . Therefore, we vary the values of β as follows:

$$\beta = \beta_0 + M - 0.269 \times k \Phi^{-1}(0.8)/4, \quad k = 2, 3, \dots, 12.$$

The null values of β for testing $H_{0b}: P(\beta - \beta_0 - M) = 0.8$ corresponding to $\sigma = k \times 0.1345$, $k = 1, \dots, 4$ are 0.0368, -0.0764, -0.1896, and -0.3028, respectively. All simulation results are based on 10,000 replicates.

To examine the impact of the total sample size on the power of the test, we consider the case when $\beta = \beta_0 + M - 2.69 \times \Phi^{-1}(0.8)/4$, which yields reasonable powers for comparison. Figure 2 displays the powers of the signed LRT using the REML estimators in this case. As expected, the power increases as n increases. However, the rate of the increment of the power is small especially when σ is large.

In the second set of simulation studies, we examine the performance of the proposed allocation rule and compare it with the allocation rule in the ASPECCT study and the equal allocation rule. We consider the same simulation setting as above. Figure 3 presents the type I error rates and powers of the signed LRT using the REML estimators for testing $H_{0b}: P(\beta_l - \beta_0 - M) = 0.7$. In all cases, the proposed allocation rule performs as well as or outperforms the equal allocation rule and the original allocation rule in the ASPECCT study.

4. The ASPECCT study

We apply the proposed signed LRTs to the ASPECCT study. Four regions with five or less patients were excluded from the analysis. Table 1 presents the data from this study including the log-hazard ratio estimates $\hat{\beta}_l$, standard error estimates $\hat{\sigma}_l$, sample sizes n_l , number of events d_l and $\hat{v}_l$, $l = 1, \dots, 23$. We note that the country level data for the ASPECCT study has been modified slightly since the study is still ongoing.

We analyze the data and obtain the MLEs and REML estimators of β and σ at $(\hat{\beta}, \hat{\sigma}) = (-0.058, 0.269)$ and $(\hat{\beta}_{REML}, \hat{\sigma}_{REML}) = (-0.058, 0.254)$. We apply the proposed signed LRT tests to the real data and consider four scenarios for the margin, $M = 0.151, 0.201, 0.251$, and 0.301 . We also vary p from 0.5 to 0.90 . Table 2 presents the p -values of the tests using the MLEs and the REML estimators. For $M = 0.251$, no significant results were detected for either method except when $p = 0.5$, under which the tests are equivalent to testing the overall noninferior effect of the drug because $\Phi^{-1}(0.5) = 0$. When $M = 0.301$, there is strong evidence that at least 70% of the regions have noninferior treatment effects. Note that the MLE of σ is negatively biased leading to a smaller estimator of γ compared to γ_{REML} and therefore, the signed LRT using MLEs tends to have smaller p -values than the signed LRT using the REML estimators.

We also perform a residual analysis to check the goodness-of-fit of the model to the ASPECCT data. We calculate the fitted values of the random effects as:

$$\hat{u}_l = \frac{\hat{\sigma}_{REML}^2}{\hat{\sigma}_{REML}^2 + \hat{\sigma}_l^2} (\hat{\beta}_l - \hat{\beta}_{REML})$$

and:

$$\hat{\varepsilon}_l = \frac{\hat{\sigma}_l^2}{\hat{\sigma}_{REML}^2 + \hat{\sigma}_l^2} (\hat{\beta}_l - \hat{\beta}_{REML}).$$

Define $\hat{W}_k = \frac{1}{\hat{\sigma}_{REML}^2 + \hat{\sigma}_k^2} / \sum_{l=1}^L \frac{1}{\hat{\sigma}_{REML}^2 + \hat{\sigma}_l^2}$. It can be shown that $\hat{\beta}_{REML} = \sum_{l=1}^L \hat{W}_l \hat{\beta}_l$. Therefore, the standardized predictors of the random effects (u_l 's) are given by:

$$\hat{u}_l^* = (\hat{\beta}_l - \hat{\beta}_{REML}) / \sqrt{\hat{h}_l},$$

where $\hat{h}_l = \sum_{k \neq l}^L \hat{W}_k^2 (\hat{\sigma}_{REML}^2 + \hat{\sigma}_k^2) + (1 - \hat{W}_l)^2 (\hat{\sigma}_{REML}^2 + \hat{\sigma}_l^2)$. Figure 4 presents the histograms and normal Q-Q plots of $\hat{u}_l^*$. The observed data points are close to a straight line with intercept 0 and slope 1, suggesting a good fit of the model to the data.

5. Discussion

In this paper, we propose an intuitive definition of consistency quantified by the percentage of regions with noninferior treatment effects. Under this definition, the treatment effects are consistent if a prespecified percentage of regions have noninferior treatment effects even if the treatment effects are very different. It is worth to note that it is subjective to select the threshold and future discussion on this topic is needed. Alternatively, we can define consistency such that the treatment effects across regions are similar to each other. This definition requires that the treatment effects should not differ much from region to region. For this consistency definition, one quantity for assessing the variability of the treatment effects across regions is the so-called coefficient of variation $CV = \sigma / |\beta - \beta_0|$, which describes how different region effects vary with respect to the whole population average deviation from β_0 . This quantity has been popularly used in assessing measurement errors in biomarkers. Therefore, if CV is less than a prespecified threshold c_0 (say, 20%), we would then conclude that consistency is obtained. In other words, we will declare success of an MRCT if $\beta - \beta_0 \geq M$ and $CV \leq c_0$, i.e., $\beta - \beta_0 \geq M$ and $|\beta - \beta_0| > \sigma / c_0$. This implies that the null hypothesis is:

$$H_{0c}^*: \beta - \beta_0 > M \text{ or } |\beta - \beta_0| \leq \sigma / c_0.$$

Let $\hat{\beta}$ and $\hat{\sigma}$ be the MLEs or the REMLs estimator from model (2). Then we reject H_{0c} if:

$$\hat{\beta} - \beta_0 - M \leq c_1 \text{ and } |\hat{\beta} - \beta| / \hat{\sigma} \geq c_2$$

for some constants c_1 and c_2 which will be determined by joint distribution of $(\hat{\beta}, \hat{\sigma})$ under H_{0c}^* . As a referee pointed out, since the threshold c_0 has some subjectivity, then, similar to the overall noninferiority margin, the threshold should also be discussed with regulators.

Based on the definition of consistency in Section 2, consistent noninferior treatment effects across regions imply global noninferior treatment effects. Therefore, one may first test whether there is significant global noninferior treatment effect and then test for the consistency of noninferior treatment effects across regions given that the global noninferior treatment effect is significant. To control the overall type I error of the sequential test, we determine critical values c_1 and c_2 such that:

$$\mathbf{P}(\hat{\beta} - \beta_0 - M_1 < c_1 | H_{0g}) + \mathbf{P}(\hat{\gamma} - \beta_0 - M_2 < c_2 | \hat{\beta} - \beta_0 - M_1 < c_1, H_{0c}) = \alpha.$$

where M_1 and M_2 are the margins for testing the global noninferior drug effect and the consistency of noninferior treatment effects across regions, respectively. Note that we can

use more stringent margins for assessing a global treatment effect compared to assessing the consistency of treatment effects across regions. This methodology is the subject of future work.

In this paper, we have used a unified margin to assess the global consistency based on the definition in Section 2. Research is underway to expand this method to handle region-specific consistency margins (in this case M_I for $I = 1, \dots, L$). However, the use of region-specific margins for global consistency should be performed with caution in MRCTs (Tsong and Tsou, 2013). Defining the region-specific margins per FDA Guidelines (FDA, 2016) requires information on the magnitude and variability of the estimated control treatment effect (over placebo) in each of the regions. As discussed in Section 2, this information is usually unavailable hence the argument for a unified margin (which is based on the always available historical global control effect). Region-specific margins can be justified when local regulatory agencies require the assessment of regional consistency. It should be noted that regional consistency is more meaningful when positive conclusions on global noninferiority and global consistency have been reached. How to conduct these analyses with either a unified or a sequential approach is also the subject of future work.

This paper focused on noninferiority trials. The goal of noninferiority trial is not only to show that the new drug is noninferior to the active control, but also to show more importantly that the new drug is better than the placebo (FDA, 2016). Therefore for noninferiority trials, evaluating treatment effects (to ensure the new drug is better than the placebo) in all regions is fundamentally necessary compared with superiority trials (even if sample sizes may be small for some regions). Note that in the reality of drug development, the testing for regional consistency (i.e., are treatment effects consistent across all regions?) is different from the testing for overall noninferiority (i.e., is the new treatment non-inferior to the active control?); therefore could warrant different noninferiority margins. Based on our experience, noninferiority margin for regional consistency can be much less tight depending on the active control effect on placebo and sample sizes for each region. With a relaxed margin, the key for evaluating regional consistency is not whether treatment effects are positive, rather the treatment effects are above a predefined threshold (hence consistent) in most if not all regions.

Acknowledgments

The authors wish to thank the associate editor and one anonymous referee for their helpful comments that have greatly improved the presentation of this paper. The authors also thank Amgen's Department of Global Biostatistics for providing the data on the ASPECTT trial and Dr. Chunlei Ke at Amgen for helpful discussions.

References

- Barndorff-Nielsen OE. Infereni on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*. 1986; 73:307–322.
- Chen J, Quan H, Binkowitz B, Ouyang SP, Tanaka Y, Li G, Menjoge S, Ibia E. Assessing consistent treatment effect in a multi-regional clinical trial: A systematic review. *Pharmaceutical Statistics*. 2010; 9:242–253. [PubMed: 20872623]
- Committee for Medicinal Products for Human Use (CHMP). Reflection paper on the extrapolation of results from clinical studies conducted outside the European Union (EU) to the EU. 2009

- FDA. Non-Inferiority Clinical Trials to Establish Effectiveness Guidance for Industry. U.S. Department of Health and Human Services. Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research; 2016.
- Hsieh, ET-C., Liu, J-p. Interaction effects in bridging studies. In: Liu, J-pChow, S-C., Hsiao, C-F., editors. Design and Analysis of Bridging Studies. Boca Raton, FL: CRC Press; 2013.
- Hung H, Wang SJ, O'Neill RT. Consideration of regional difference in design and analysis of multi-regional trials. *Pharmaceutical Statistics*. 2010; 9:173–178. [PubMed: 20872619]
- ICH E17. Final Concept Paper. General Principles on Planning/Designing Multi-Regional Clinical Trial. 2014. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E17/E17_Final_Concept_Paper_July_2014.pdf. (accessed November 2016)
- ICH, International Conference on Harmonisation. Tripartite guidance e5 ethnic factors in the acceptability of foreign data. The U.S. Federal Register. 1998; 83:31790–31796.
- Jonker DJ, O'Callaghan CJ, Karapetis CS, Zalcberg JR, Tu D, Au HJ, Berry SR, Krahn M, Price T, Simes RJ, et al. Cetuximab for the treatment of colorectal cancer. *New England Journal of Medicine*. 2007; 357:2040–2048. [PubMed: 18003960]
- Kalbfleisch, JD., Prentice, RL. The Statistical Analysis of Failure Time Data. Vol. 360. Hoboken, NJ: John Wiley & Sons; 2011.
- Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, Simes RJ, Chalchal H, Shapiro JD, Robitaille S, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine*. 2008; 359:1757–1765. [PubMed: 18946061]
- Liu, J-p, Chow, SC., Hsiao, CF. Design and Analysis of Bridging Studies. CRC Press; 2013.
- Liu JP, Lin JR, Hsieh E. A noninferiority test for treatment-by-factor interaction with application to bridging studies and global trials. *Drug Information Journal*. 2009; 43:11–16.
- Ministry of Health, Labour and Welfare of Japan. Basic Principles on Global Clinical Trials. Tokyo, Japan: Pharmaceuticals and Medical Devices Agency; 2007.
- Price TJ, Peeters M, Kim TW, Li J, Cascinu S, Ruff P, Suresh AS, Thomas A, Tjulandin S, Zhang K, et al. Panitumumab versus cetuximab in patients with chemotherapy-refractory wild-type kras exon 2 metastatic colorectal cancer (aspsect): A randomised, multicentre, open-label, non-inferiority phase 3 study. *The Lancet Oncology*. 2014; 15:569–579. [PubMed: 24739896]
- Quan H, Li M, Chen J, Gallo P, Binkowitz B, Ibia E, Tanaka Y, Ouyang SP, Luo X, Li G, et al. Assessment of consistency of treatment effects in multiregional clinical trials. *Drug Information Journal*. 2010; 44:617–632.
- Quan H, Li M, Shih WJ, Ouyang SP, Chen J, Zhang J, Zhao PL. Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials. *Statistics in Medicine*. 2013; 32:1691–1706. [PubMed: 22855311]
- Quan H, Shih WJ. Treatment by stratum interaction test for equivalence and non-inferiority trials. Proceedings of the Biopharm Section, Joint Stat. Meetings, Am Stat Assoc., CD Version, ASA. 2002
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
- Shih, WJ., Quan, H. Consistency of treatment effects in bridging studies and global multiregional trials. In: Liu, J-pChow, S-C., Hsiao, C-F., editors. Design and Analysis of Bridging Studies. Boca Raton, FL: CRC Press; 2013.
- Tsong, Y., Tsou, HH. Multiregional clinical trials. In: Liu, J-pChow, S-C., Hsiao, C-F., editors. Design and Analysis of Bridging Studies. Boca Raton, FL: CRC Press; 2013.
- Van Cutsem E, Peeters M, Siena S, Humblet Y, Hendlisz A, Neyns B, Canon JL, Van Laethem JL, Maurel J, Richardson G, et al. Open-label phase III trial of panitumumab plus best supportive care compared with best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer. *Journal of Clinical Oncology*. 2007; 25:1658–1664. [PubMed: 17470858]
- Wiens BL, Heyse JF. Testing for interaction in studies of noninferiority. *Journal of Biopharmaceutical Statistics*. 2003; 13:103–115. [PubMed: 12635906]

Figure 1.

Sizes and powers for testing H_{0b} with $N = 575$. Solid and dashed curves correspond to the sizes and powers of the signed likelihood ratio test using the MLEs and the REML estimators, respectively. Black, blue, red, and green curves correspond to the sizes and powers when $\sigma = 0.1345, 0.269, 0.4035$, and 0.538 , respectively. The null values of β for $\sigma = 0.5\hat{\sigma}, \hat{\sigma}, 1.5\hat{\sigma}$, and $2\hat{\sigma}$ are $0.0368, -0.0764, -0.1896$, and -0.3028 , respectively.

Figure 2.

Powers for testing H_{0b} with $\beta = -0.416$. —, $\sigma = 0.5\hat{\sigma}$; —, $\sigma = \hat{\sigma}$; —, $\sigma = 1.5\hat{\sigma}$; —, $\sigma = 2\hat{\sigma}$. Sample size of $N = 1150$ corresponds to the sample size in the real data.

Figure 3.
Sizes and powers for testing H_{0b} with $N = 1150$ under different allocation rules.

Figure 4.
Histogram and normal Q–Q plot of $\hat{u}^*$.

Table 1

ASPECCT data.

Region	$\hat{\beta}_l$	$\hat{\sigma}_l^2$	n_l	d_l	$\hat{v}_l$
1	0.337	0.027	110	95	2.959
2	0.066	0.549	20	20	10.982
3	0.167	0.057	110	50	6.283
4	-1.578	1.103	20	10	22.050
5	-0.751	0.225	40	25	8.987
6	-0.322	0.036	90	75	3.283
7	0.002	0.227	30	20	6.797
8	-0.209	0.087	60	40	5.221
9	0.122	0.132	50	30	6.625
10	-0.124	0.010	220	170	2.244
11	0.225	0.159	30	25	4.776
12	0.723	0.375	20	15	7.491
13	-0.316	0.280	30	20	8.395
14	-0.544	0.618	20	10	12.356
15	0.412	0.591	10	10	5.914
16	0.359	0.045	70	65	3.146
17	-0.782	0.047	90	65	4.199
18	-1.104	1.145	20	10	22.898
19	-0.326	0.931	10	10	9.312
20	0.167	0.138	30	25	4.129
21	0.513	0.256	30	20	7.681
22	-1.115	1.147	10	10	11.470
23	0.012	0.210	30	20	6.293

Table 2

p -Values for testing $H_{0b}: P(\beta_I | M) = p$.

M	p	p -value (MLE)	p -value (REML)
0.151	0.50	0.019	0.021
	0.70	0.246	0.273
	0.75	0.371	0.406
	0.80	0.516	0.556
	0.85	0.666	0.706
	0.90	0.803	0.835
0.201	0.50	0.006	0.007
	0.70	0.135	0.155
	0.75	0.229	0.259
	0.80	0.357	0.397
	0.85	0.513	0.558
	0.90	0.681	0.725
0.251	0.50	0.002	0.002
	0.70	0.068	0.081
	0.75	0.130	0.153
	0.80	0.228	0.262
	0.85	0.367	0.412
	0.90	0.545	0.595
0.301	0.50	0.001	0.001
	0.70	0.032	0.040
	0.75	0.069	0.084
	0.80	0.135	0.162
	0.85	0.246	0.285
	0.90	0.411	0.462

Note. Testing of H_{0b} for $p = 0.5$ corresponds to the testing of the overall treatment effect because $\phi^{-1}(0.5) = 0$.