Power and sample size computations for simultaneously assessing consistency of treatment effects

Andreas Kitsche[a]*

In this paper, the problem of calculating the power and sample size for the simultaneous assessment of consistency of treatment effects is addressed. The method is based on a general formulation of consistency as treatment-by-subset interaction, while the interaction term is defined as the ratio of treatment effects. This approach allows the interpretation of inconsistency as relative change of the treatment effects. In addition, conclusions are based on an appropriate defined consistency margin. The methodology is applicable in trials with a continuous as well as a binary endpoint. Two power definitions arising in multiple testing, namely the all-pair (complete) power and any-pair (minimal) power, are considered within this manuscript. While the focus of this paper is on the assessment of consistency in multi-regional clinical trials, the presented methodology is in general applicable for the assessment of treatment-by-subset interactions, including the detection of qualitative interactions. Several examples from clinical trials illustrate the application of the proposed procedure. For the analysis, the author developed the R add-on package `poco`, which provides the functionality presented here.

Keywords: multi-regional clinical trials, heterogeneity, treatment-by-subset interaction

1. Introduction

The problem of assessing consistency of the treatment effect among subsets of patients arises in a variety of situations in biomedical research. In recent years the assessment of consistency became an important issue in multi-regional clinical trials (MRCTs) because global development in pharmaceutical industry has received increasing attention. In MRCTs patients are enrolled from different countries or geographic regions under one common protocol. Many authors proposed methods for consistency assessment in MRCTs , see e.g. Quan et al. (2010a); Tsou et al. (2012); Quan et al. (2013) and Chen et al. (2013). The wide variety of presented methods is based upon a large set of definitions introduced for consistency assessment. In a review paper Quan et al. (2010a) presented and discussed a number of definitions for assessing consistency of treatment effects in a MRCT. Definition 1 in their paper defines consistency by demonstrating that regional effects exceed a proportion of the overall effect. This definition was also recommended by the Ministry of Health, Labour and Welfare (MHLW) of Japan, issued in the "Basic concepts for Joint International Clinical Trials" guidance (2007). In that document it is proposed that the observed treatment effect for Japanese patients should be at least half of that observed for all patients to accept consistency of the treatment effect. In a recently published paper, Kitsche and Hothorn (2014) introduced a method to detect treatment-by-subset interaction, while the interaction term is defined as the ratio of treatment effects. As discussed by Kitsche and Hothorn their approach is in general applicable for the above mentioned definition of consistency in MRCTs. Their methodology allows the interpretation of consistency as relative change of a

[a] *Institut für Biostatistik, Leibniz Universität Hannover, Herrenhäuser Strasse 2, 30419 Hannover, Germany*
* *Correspondence to: Institut für Biostatistik, Leibniz Universität Hannover, Herrenhäuser Strasse 2, 30419 Hannover, Germany. E-mail: kitsche@biostat.uni-hannover.de*

Contract/grant sponsor:

subset specific treatment effect in contrast to a reference treatment effect, e.g. the overall treatment effect. In addition conclusions are based on an appropriate defined consistency margin, that allows to set up a region or subset specific definition of consistency. Besides of its application for quantitative response variables, their approach is also applicable in the presence of binary response variables, as demonstrated by Kitsche (2014).

Quan et al. (2010a) briefly presented formulas for calculating the probability for assessing consistency at the design stage. Li et al. (2012) provide the corresponding R functions (R Core Team, 2013) for calculating the unconditional and conditional probabilities for demonstrating consistency in relation with the overall/regional sample sizes. Ikeda and Bretz (2010) derived a closed form expression for the probabilities based on the method proposed by the Japanese regulatory guideline. Additionally, they proposed an alternative method with better operating characteristics. Nevertheless, none of the published methods considers sample size and power calculations for the simultaneous assessment of consistency for each region. Therefore, the focus of this paper is on calculating the probability for the simultaneous assessment of consistency defined as a pre-specified relative change from the overall treatment effect. Two power definitions arising in multiple hypothesis testing are considered: the all-pair (complete) power and the any-pair (minimal) power (Horn and Vollandt, 1998). The any-pair power is defined as the probability to detect at least one true inconsistent treatment effect, whereas the all-pair power is given as the probability to find all true inconsistent treatment effects (Dilba et al., 2006).

The paper is organized as follows. The models for the continuous and the binomial response variable in a two-way layout are introduced in Section 2. Afterwards, the hypotheses for the assessment of consistency among subsets of patients are formulated in the context of MRCTs. The formulas for the power and sample size calculations are presented in Section 2.5. Several examples from clinical trials are evaluated in Section 3 to illustrate the broad application of the methodology. To conclude the paper a brief discussion is given in Section 4.

2. Methods

2.1. Continuous response variable

At first, the model for a normally distributed outcome measure is introduced. Afterwards the method is extended to a binary response variable in Section 2.2. We consider a completely randomized two-way layout where the primary endpoint is a continuous and normally distributed outcome measure. The corresponding cell means model is given by

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \tag{1}$$

where the index i (with $i = 1, \ldots, I$) denotes the level of the treatment factor, and j (with $j = 1, \ldots, J$) denotes the level of a subsetting factor, e.g., regions in a MRCT, centres in a multi-centre trial. For the sake of convenience and without loss of generality, we set the number of groups to two, representing a placebo group ($i = 1$) and one active treatment group ($i = 2$). The number of independent replicates for the ijth treatment-by-subset combination is given by n_{ij} with the index k (with $k = 1, \ldots, n_{ij}$). The error effect associated with the independent observation on unit k is assumed to be normally distributed around zero with a common variance, $\epsilon_{ijk} \sim N(0, \sigma^2)$. The total sample size is given by $N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$. The parameter vector of cell means is presented by $\boldsymbol{\mu} = (\mu_{11}, \mu_{21}, \ldots, \mu_{1J}, \mu_{2J})$, where the elements of $\boldsymbol{\mu}$ are primarily ordered according to the subsetting factor and within the subsets according to the treatment factor. For the sake of simplicity, we give the vector $\boldsymbol{\mu}$ the index l, resulting in $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)$, where $L = I \cdot J$, the number of treatment-by-subset combinations. The vector of estimates is given by $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_L)$, where $\hat{\mu}_l = \bar{y}_l = \sum_{k=1}^{n_l} y_{lk}/n_l$. The pooled sample variance is defined by $s^2 = \sum_{l=1}^{L} \sum_{k=1}^{n_l} (y_{lk} - \bar{y}_l)^2 / (N - L)$. Furthermore, a treatment effect is defined as the difference between the active treatment group and the placebo group for each subset, $\delta_j = \mu_{2j} - \mu_{1j}$.

2.2. Binary response variable

Within this subsection we suppose that the primary endpoint Y is a binary outcome represented by 1 and 0, with generic labels success and failure, having success probability of π_l. It is assumed that the total number of successes in each factor combination is given by y_l, whereas y_l follows a binomial distribution with parameters n_l and π_l by $bin\,(n_l, \pi_l)$. The success probability π_l is defined as the probability of observing an event in the lth factor combination, $\pi_l = P\,(Y = 1 | l = L)$. The corresponding vector of maximum likelihood estimators for the sample proportions is given by $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_L)$, with $\hat{\pi}_l = y_l/n_l$. The related standard error is defined by $\hat{\sigma}(\pi_l) = \sqrt{\pi_l(1 - \pi_l)/n_l}$ (Agresti and Ryu, 2010). Again, a treatment effect is defined as the difference of the success probability between the two treatment groups, which is commonly denoted as risk difference $\delta_j = \pi_{2j} - \pi_{1j}$.

2.3. Hypotheses for consistency assessment

As mentioned in Section 1 several definitions of consistency were considered in the literature, see e.g. Quan et al. (2010a). In this section the global hypothesis to detect an inconsistent treatment effect is formulated according to Method 1 in the Japanese guideline and Definition 1 proposed by Quan et al. (2010a). They declared consistency when achieving in each region a specified proportion of the overall effect:

$$\delta_j/\delta \geq \theta, \tag{2}$$

where δ_j is the treatment effect for the jth region, δ defines the overall treatment effect and θ denotes the relative consistency margin.

It is straightforward to formulate the definition in Eq. 2 simultaneously for all regions using the methodology introduced by Kitsche and Hothorn (2014), which was originally proposed to assess qualitative interactions in clinical trials. They demonstrate that the hypothesis in Eq. 2 can be formulated as ratio of user-defined linear combinations of treatment means $\gamma_j = \boldsymbol{h}_j\boldsymbol{\mu}/\boldsymbol{d}_j\boldsymbol{\mu}$, where γ_j defines the ratio parameter. Following the notation of Kitsche and Hothorn (2014) the local null and alternative hypotheses for the jth region are given by

$$H_0^j: \quad \frac{\boldsymbol{h}_j\boldsymbol{\mu}}{\boldsymbol{d}_j\boldsymbol{\mu}} \geq \theta \qquad H_A^j: \quad \frac{\boldsymbol{h}_j\boldsymbol{\mu}}{\boldsymbol{d}_j\boldsymbol{\mu}} < \theta, \tag{3}$$

where $\boldsymbol{h}_j$ and $\boldsymbol{d}_j$ denote the jth numerator and denominator contrasts. A contrast or a comparison among means is defined as a linear combination among the means that have known coefficients h_{jl} and d_{jl} for the jth region and the lth treatment-by-subset combination. To formulate the global null and alternative hypotheses in a unified framework for all regions the jth numerator and denominator contrasts can be summarized in the $J \times L$ numerator and denominator contrast matrix $\boldsymbol{C}_{\text{Numerator}} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_J)$ and $\boldsymbol{C}_{\text{Denominator}} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_J)$. The general form of the contrast matrices $\boldsymbol{C}_{\text{Numerator}}$ and $\boldsymbol{C}_{\text{Denominator}}$ for the region specific comparisons of the regional treatment effect to the overall treatment effect is given as (Kitsche and Hothorn, 2014):

$$\boldsymbol{C}_{\text{Numerator}} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & 1 & -1 & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots & 1 & -1 \end{pmatrix} \tag{4}$$

and

$$\boldsymbol{C}_{\text{Denominator}} = \begin{pmatrix} \frac{n_1}{N} & -\frac{n_1}{N} & \frac{n_2}{N} & -\frac{n_2}{N} & \cdots & \cdots & \frac{n_J}{N} & -\frac{n_J}{N} \\ \frac{n_1}{N} & -\frac{n_1}{N} & \frac{n_2}{N} & -\frac{n_2}{N} & \cdots & \cdots & \frac{n_J}{N} & -\frac{n_J}{N} \\ \vdots & \vdots & \vdots & \vdots & \frac{n_j}{N} & -\frac{n_j}{N} & \vdots & \vdots \\ \frac{n_1}{N} & -\frac{n_1}{N} & \frac{n_2}{N} & -\frac{n_2}{N} & \cdots & \cdots & \frac{n_J}{N} & -\frac{n_J}{N} \end{pmatrix}. \tag{5}$$

Obviously, each row in $\boldsymbol{C}_{\text{Numerator}}$ defines the linear combination of treatment means that result in the regional treatment effect δ_j, and each row in $\boldsymbol{C}_{\text{Denominator}}$ defines the linear combination of treatment means that corresponds to the overall treatment effect δ.

2.4. Simultaneous test for consistency assessment

Kitsche and Hothorn (2014) derived the test statistic associated with the hypothesis in Eq. 3 by a reformulation of the ratio problem as a linear form $L_j = (\boldsymbol{h}_j - \theta\boldsymbol{d}_j)\,\boldsymbol{\mu}$. A single test statistic to test the jth local null hypothesis in Eq. 3 is given by the ratio of the linear form to its appropriate standard error:

$$T_j = \frac{(\boldsymbol{h}_j - \theta\boldsymbol{d}_j)\,\hat{\boldsymbol{\mu}}}{S\sqrt{(\boldsymbol{h}_j - \theta\boldsymbol{d}_j)\,\boldsymbol{M}\,(\boldsymbol{h}_j - \theta\boldsymbol{d}_j)}}, \tag{6}$$

where $\boldsymbol{M}$ is a diagonal matrix including the reciprocals of the sample sizes n_l, and S^2 is the pooled variance estimator of the common variance σ^2 based on $\nu = \sum_{l=1}^{L} = (n_l - 1)$ degrees of freedom. Under the global null hypothesis the vector of test statistics $\boldsymbol{T} = (T_1, \ldots, T_J)$ jointly follows a central J-variate t-distribution with ν degrees of freedom and a correlation matrix $\boldsymbol{R} = [\varsigma_{jj'}]$. Each element of the correlation matrix $\boldsymbol{R}$ is given by

$$\varsigma_{jj'} = \frac{(\boldsymbol{h}_j - \theta\boldsymbol{d}_j)\,\boldsymbol{M}\,(\boldsymbol{h}_{j'} - \theta\boldsymbol{d}_{j'})}{\sqrt{(\boldsymbol{h}_j - \theta\boldsymbol{d}_j)\,\boldsymbol{M}\,(\boldsymbol{h}_j - \theta\boldsymbol{d}_j)}\sqrt{(\boldsymbol{h}_{j'} - \theta\boldsymbol{d}_{j'})\,\boldsymbol{M}\,(\boldsymbol{h}_{j'} - \theta\boldsymbol{d}_{j'})}}.$$

The one-sided null hypothesis in Eq. 3 is rejected if $T_j < t_{\alpha,\nu,\boldsymbol{R}}$, where $t_{\alpha,\nu,\boldsymbol{R}}$ denotes the α-level equi-coordinate percentage point from the multivariate t-distribution $Mt_{\nu,\boldsymbol{R}}$.

2.5. Power and sample size computations

In this section, the power associated with the test described in the previous section is provided. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$ denote a vector of ratios on the basis of which we compute the power. Under the alternative hypothesis some γ_j may be less than θ. When some of the H_A^j's are true, the vector of test statistics $\boldsymbol{T}$ follows a non-central J-variate t-distribution $Mt_{\nu,\boldsymbol{R},\boldsymbol{\tau}}$ with ν degrees of freedom, correlation matrix $\boldsymbol{R}$, and non-centrality parameter vector $\boldsymbol{\tau}$. The elements of the non-centrality parameters are given by

$$\tau_j = \frac{(\boldsymbol{h}_j - \theta \boldsymbol{d}_j)\,\boldsymbol{\mu}}{\sigma\sqrt{(\boldsymbol{h}_j - \theta \boldsymbol{d}_j)\,\boldsymbol{M}\,(\boldsymbol{h}_j - \theta \boldsymbol{d}_j)}}.$$

In the following it is demonstrated that the presented non-centrality parameter is a generalization of the non-centrality parameter proposed by Hauschke and Kieser (2001) and Dilba et al. (2006) for the assessment of non-inferiority. Dilba et al. (2006) presented a methodology to calculate the power and sample sizes associated with simultaneous test for non-inferiority in the case of comparing several experimental treatments with an active control. The hypotheses that correspond to a test for non-inferiority can be formulated by using the general formulation given in Eq. 3. In this situation the vector of cell means $\boldsymbol{\mu}$ simplifies to the vector of means of the active and experimental treatment group. Furthermore, the numerator and denominator contrast matrices are given by:

$$\boldsymbol{C}_{\text{Numerator}} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{C}_{\text{Denominator}} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 0 \end{pmatrix}.$$

As a consequence the non-centrality parameter vector is given by

$$\begin{aligned} \tau_j &= \frac{(\boldsymbol{h}_j - \theta \boldsymbol{d}_j)\,\boldsymbol{\mu}}{\sigma\sqrt{(\boldsymbol{h}_j - \theta \boldsymbol{d}_j)\,\boldsymbol{M}\,(\boldsymbol{h}_j - \theta \boldsymbol{d}_j)}} \\ &= \frac{\mu_l - \theta\mu_1}{\sigma\sqrt{\frac{1}{n_1} + \frac{\theta^2}{n_l}}} \\ &= \frac{\frac{\mu_l}{\mu_1} - \theta}{\kappa_0\sqrt{\frac{1}{n_l} + \frac{\theta^2}{n_0}}} \end{aligned} \tag{7}$$

where $\kappa_0 = \sigma/\mu_1$ denotes the coefficient of variation of the control group (this corresponds to the Eq. (11) in Hauschke et al. (1999)).

The numerical computation routines presented in Bretz et al. (2001) can be used for the computation of non-central multivariate t-probabilities. Furthermore, γ^* denotes the smallest irrelevant ratio parameter which is to be detected and $I(\gamma^*)$ defines the set of indices $I(\gamma^*) = \{j|\gamma_j < \gamma^*\} = \{j_1, \ldots, j_m\}$,$1 \leq m \leq J$. All regions with ratio parameters smaller than γ^* are inconsistent to the overall treatment effect and therefore under the alternative hypothesis.

Any-pair power Suppose the goal is now to detect at least one true inconsistent treatment effect according to the definition presented in Section 2.3 (Eq. 3) with a given power of $1 - \beta$. The power associated with this problem is called the any-pair (minimal) power and is given by

$$\begin{aligned} Q_{\text{any-pair}} &= P\left(\bigcup_{j \in I(\gamma^*)} T_j < t_{\alpha,\nu,\boldsymbol{R},\boldsymbol{\tau}}\right) \\ &= 1 - P\left(T_j \geq t_{\alpha,\nu,\boldsymbol{R},\boldsymbol{\tau}}, \exists j \in I(\gamma^*)\right) \end{aligned} \tag{8}$$

where $t_{\alpha,\nu,\boldsymbol{R},\boldsymbol{\tau}}$ is the α quantile of the non-central multivariate t-distribution $Mt_{\nu,\boldsymbol{R},\boldsymbol{\tau}}$.

All-pairs power Suppose interest is to detect all true inconsistent treatment effects with a given power $1-\beta$. This is called the all-pairs (complete) power. The power of this test is given by

$$Q_{\text{all-pairs}} = P\left(\bigcap_{j\in I(\gamma^*)} T_j < t_{\alpha,\nu,\boldsymbol{R},\boldsymbol{\tau}}\right) \tag{9}$$

$$= P\left(T_j < t_{\alpha,\nu,\boldsymbol{R},\boldsymbol{\tau}}, \forall j \in I(\gamma^*)\right).$$

The main objective in practical design studies is the determination of the required sample size for a specific trial. For given values of α, β, vector of cell means $\boldsymbol{\mu}$, vector of consistency margins $\boldsymbol{\theta}$ and common variance σ^2 the goal is to compute the vector of sample sizes $\boldsymbol{n} = (n_1, \ldots, n_L)$. This problem can be solved by using the above power expressions via an iterative procedure. For simplicity, a balanced design with n observations per treatment-by-subgroup combination is considered. The required sample size n is determined iteratively by starting with a given sample size, evaluating the corresponding power and a decision upon the results to see whether a higher sample size is required or not. This procedure is repeated subsequently until the power condition $Q \geq 1-\beta$ is satisfied.

Software The add -on package `mvtnorm` (Genz and Bretz, 1999) of the open source software R (R Core Team, 2013) provides the functionality for the calculations from the multivariate normal- and t-probabilities. The add-on package `poco` with its functions `PowCon` and `nPowCon` calculates the power and sample size for the ratios of treatment differences for the all-pair and any-pair power definition. In addition, the package `poco` contains a vignette for a detailed documentation of its functionalities and several example data sets with some guidance for their post-hoc power analysis.

To install the package `poco` directly from `github`, the package `devtools` is needed:

```
#install.packages("devtools")
library(devtools)
install_github("AKitsche/poco")
library(poco)
```

Up to this point the power and sample size formulas were presented for continuous variables. Appendix A provides the formulas for the any-pair and all-pair power definition in the binomial case.

3. Illustrative examples

In this section the proposed methodology for calculating the power and sample size to detect an inconsistent treatment effect is demonstrated on several clinical trial examples. Since the presented studies are all enclosed, a post-hoc power analysis is conducted to assess the power of the study given the observed effects. For the post-hoc power assessment of the presented examples either Eq. 8 or 9 is used and it is conditioned on the estimates for σ and $\boldsymbol{\mu}$, or π, based on the observed data. In practice, it is recommended to use the presented formulas for an a-priori sample size calculation with a predefined power specification.

3.1. MERIT-HF study, two-arm, multi-regional study

The first example describes a MRCT, namely the Metoprolol Controlled-Release Randomized Intervention Trial in Heart Failure (MERIT-HF) (MERIT-HF Study Group, 1999). The large scale randomized, double blind, placebo controlled trial was conducted to investigate the treatment effect of adding once-daily doses of metoprolol controlled-release/extended-release (Meto CR/XL) to the optimum standard therapy in terms of lowering mortality in patients with symptomatic heart failure. A total number of 3991 patients were randomized into the placebo or the Meto CR/XL group in 14 countries. According to Quan et al. (2013), the data from Finland were combined with the data from Denmark, and the data from the Netherlands were combined with the data from Switzerland because no event was observed in the Meto CR/XL group in Finland and Switzerland. From Table 1, a decreasing overall treatment effect in terms of a risk difference is observable, whereas in two regions, Iceland and USA, the treatment effect increases.

The observed reversal treatment effect in the US population of the MERIT-HF trial was already part of a serious discussion in the scientific literature, see, e.g. Wittes (2013). The goal is now to calculate the post-hoc any-pair power to detect an inconsistent treatment effect given the estimated success probabilities of this study. Figure 1 presents the any-pair power depending on the consistency margin θ for a range between 0 and 1. It should be noted, that a value of $\theta = 0$ corresponds to the special case of detecting a qualitative interaction (Kitsche and Hothorn, 2014) and a value of $\theta = 1$ corresponds to the case of detecting a treatment effect that is smaller than the overall treatment effect. Given the

Table 1. Number of successes, failures, sample sizes and estimated success probabilities for each treatment-by-region combination in the MERIT-HF study.

Region (j)	Treatment	Outcome		Total	Proportion
		Success	Failure	(n_{ij})	Success
Belgium	Meto CR/XL	3	65	68	0.04
	Placebo	13	53	66	0.20
Czech Republic	Meto CR/XL	9	114	123	0.07
	Placebo	17	107	124	0.14
Denmark/Finland	Meto CR/XL	11	150	161	0.07
	Placebo	13	151	164	0.08
Germany	Meto CR/XL	19	233	252	0.08
	Placebo	31	216	247	0.13
Hungary	Meto CR/XL	16	195	211	0.08
	Placebo	29	183	212	0.14
Iceland	Meto CR/XL	2	17	19	0.11
	Placebo	2	20	22	0.09
Norway	Meto CR/XL	6	91	97	0.06
	Placebo	11	94	105	0.10
Poland	Meto CR/XL	8	94	102	0.08
	Placebo	8	94	102	0.08
Sweden	Meto CR/XL	2	37	39	0.05
	Placebo	9	37	46	0.20
The Netherland/Switzerlnd	Meto CR/XL	14	285	299	0.05
	Placebo	26	265	291	0.09
UK	Meto CR/XL	4	83	87	0.05
	Placebo	9	74	83	0.11
USA	Meto CR/XL	51	481	532	0.10
	Placebo	49	490	539	0.09
Total	Meto CR/XL	145	1845	1990	0.07
	Placebo	217	1784	2001	0.12

estimates in Table 1 and the original sample size the power to detect a regional treatment effect that is smaller than the overall treatment effect ($\theta = 1$) is 0.58. The power to detect an inconsistent treatment effect in at least one region rapidly decreases with decreasing θ. Including 50% more patients in the study would result in an any-pair power to detect a smaller regional treatment effect of 0.79 (see Figure 1).

3.2. Multi-centre clinical trial

The next example considers a multi-centre clinical trial published by Dmitrienko et al. (2005). In the multi-centre depression trial, two groups of patients, one treatment and one placebo group, were compared. The primary endpoint was the change from the baseline to the end of the nine week acute treatment phase in the 17-item Hamilton depression rating scale total score (HAMD17 score). The change of scores range from -2 to 28, and therefore, it is assume that this endpoint is approximately normally distributed. The experiment was conducted at five centres.

Applying the Gail and Simon test (Gail and Simon, 1985) and the method proposed by Kitsche and Hothorn (2014) result in a significant qualitative interaction due to a reversal treatment effect in at least one centre. The following post-hoc power analysis in performed for several choices of θ given the estimated cell means and common variance of the study. Figure 2 displays the power to detect an inconsistent treatment effect in at least one centre as a function of θ. The any-pair power to detect a qualitative interaction ($\theta = 0$) considering the original sample size of the study is 0.67. The probability to detect a reversal treatment effect with a magnitude of at least 50% ($\theta = -0.5$) is 0.07. Nevertheless, a choice of $\theta < 0$ is of less practical use and is only added for illustrative purposes. The any-pair power to detect a treatment effect that is smaller than the overall treatment effect ($\theta = 1$) is 0.99.

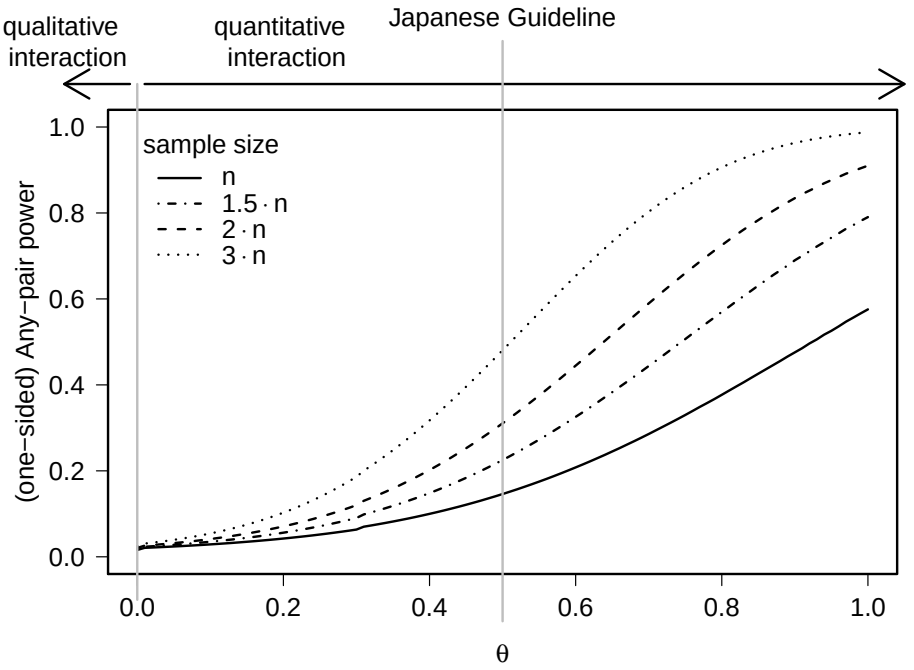

Figure 1. Any-pair power depending on the consistency margin θ conditioning on the estimates in Table 1. Several sample sizes were considered: n-original sample size;1.5n - including 50% more patients in the study; 2n - twice as the original sample size; 3n -original sample size increased threefold.

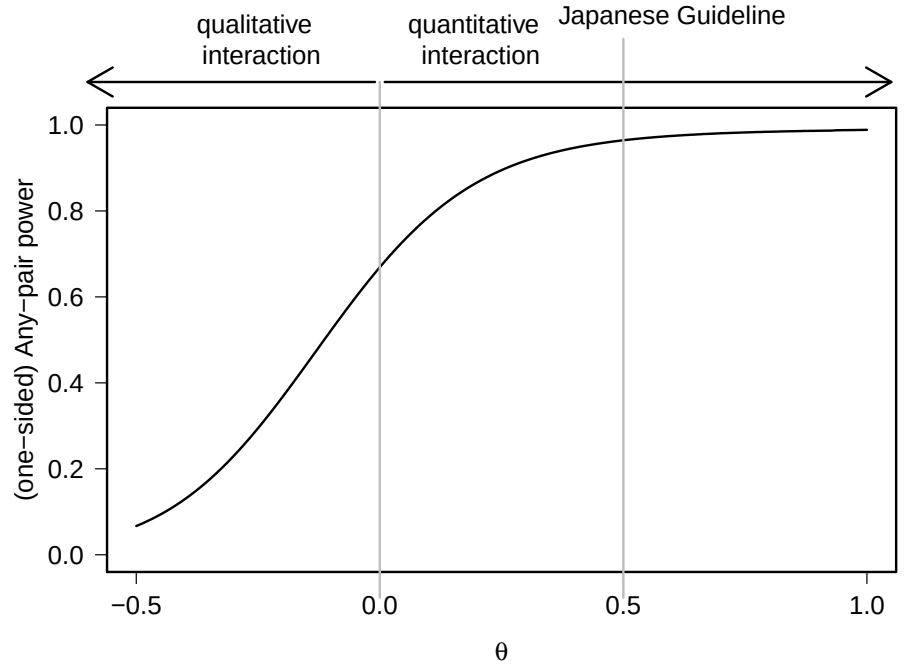

Figure 2. Power to detect at least one inconsistent treatment effect among the five centres in the Depression trial depending on the consistency margin θ

3.3. BLISS study, multi-arm, multi-regional study

The last example originates from a randomized, placebo-controlled, phase 3 trial in which belimumab (trade name Benlysta®) was compared to placebo in patients with systemic lupus erythematosus (SLE) (Navarra et al., 2011). The primary endpoint was a binary outcome measured at 52 weeks with success defined by an SLE Responder Index (SRI). Patients were enrolled from three regions, namely Asia, South America, and Eastern Europe. The trial included one placebo group and two active treatment groups, specified as low dose belimumab (1 mg/kg) and high dose belimumab (10

Table 2. Summary table with the number of successes and failures, the total number of observations, and the proportions of successes for each treatment-by-region combination in the BLISS data set. In addition, the parameters of interest $\gamma_i = \boldsymbol{h}_j\boldsymbol{\pi}/\boldsymbol{d}_j\boldsymbol{\pi}$ given $\boldsymbol{C}_{\text{Numerator}}$ and $\boldsymbol{C}_{\text{Denominator}}$ in Eq. 10 and 11 are presented..

Region (j)	Treatment	Outcome		Total	Proportion	ratio of treatment effects
		Success	Failure	(n_{ij})	Success	γ_j
Asia Pacific	Placebo	42	67	109	0.385	
	1 mg/kg belimumab	42	69	111	0.378	-0.087
	10 mg/kg belimumab	59	60	119	0.496	0.769
Eastern Europe	Placebo	71	74	145	0.364	
	1 mg/kg belimumab	85	58	143	0.618	3.198
	10 mg/kg belimumab	85	55	140	0.742	2.633
Latin America	Placebo	12	21	33	0.489	
	1 mg/kg belimumab	21	13	34	0.594	1.391
	10 mg/kg belimumab	23	8	31	0.607	0.818

mg/kg). A summary of the data set is given in Table 2.

As reported by Wittes (2013) also the FDA discussed numerical differences between the treatment effects in the pre-specified regions. In the following a post-hoc power analysis is conducted given the estimated success probabilities from Table 2. Because the treatment factor consists of three factor-levels the model presented in Section 2.2 has to be extended to $i = 1, 2, 3$. Additionally, the numerator and denominator contrast matrices have to be defined in a meaningful way, that takes the structure of the treatment factor into account. Since the primary treatment factor included one placebo group and two dose groups a Dunnett-type (many-to-one) comparison for this factor is considered here, see e.g. Bretz and Hothorn (2002) for more examples for contrasts in a one-way layout. The resulting numerator and denominator contrast matrices are given in Eq. 10 and 11.

$$\boldsymbol{C}_{\text{Numerator}} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix} \tag{10}$$

and

$$\boldsymbol{C}_{\text{Denominator}} = \begin{pmatrix} -\frac{220}{570} & \frac{220}{570} & 0 & -\frac{67}{570} & \frac{67}{570} & 0 & -\frac{283}{570} & \frac{283}{570} & 0 \\ -\frac{228}{578} & 0 & \frac{228}{578} & -\frac{65}{578} & 0 & \frac{65}{578} & -\frac{285}{578} & 0 & \frac{285}{578} \\ -\frac{220}{570} & \frac{220}{570} & 0 & -\frac{67}{570} & \frac{67}{570} & 0 & -\frac{283}{570} & \frac{283}{570} & 0 \\ -\frac{228}{578} & 0 & \frac{228}{578} & -\frac{65}{578} & 0 & \frac{65}{578} & -\frac{285}{578} & 0 & \frac{285}{578} \\ -\frac{220}{570} & \frac{220}{570} & 0 & -\frac{67}{570} & \frac{67}{570} & 0 & -\frac{283}{570} & \frac{283}{570} & 0 \\ -\frac{228}{578} & 0 & \frac{228}{578} & -\frac{65}{578} & 0 & \frac{65}{578} & -\frac{285}{578} & 0 & \frac{285}{578} \end{pmatrix}, \tag{11}$$

where each row in $\boldsymbol{C}_{\text{Denominator}}$ includes the sample sizes of the corresponding treatment-by-region combination defined by the many-to-one comparisons.

Figure 3 displays the any-pair power to detect an inconsistent treatment effect, where the treatment effects are defined as $\delta_{1j} = \pi_{2j} - \pi_{1j}$ and $\delta_{2j} = \pi_{3j} - \pi_{1j}$. Different sample sizes and consistency margins were selected to illustrate the power behaviour. In this analysis the hypotheses were formulated in a two-sided way: $H_A^{1j} : \delta_{1j}/\delta \neq \theta$ to detect an inconsistent treatment effect that is not equal to θ. As expected, the power is minimal at $\theta = 1$ because this margin means the equality of the treatment effects and therefore the majority of the parameters of interest are under the null hypothesis. The power increases with increasing distance though $\theta = 1$, since the number of parameters under the alternative hypothesis increases. Furthermore, as the sample size for each treatment-by-region combination increases, the power to detect a heterogeneous treatment effect also increases.

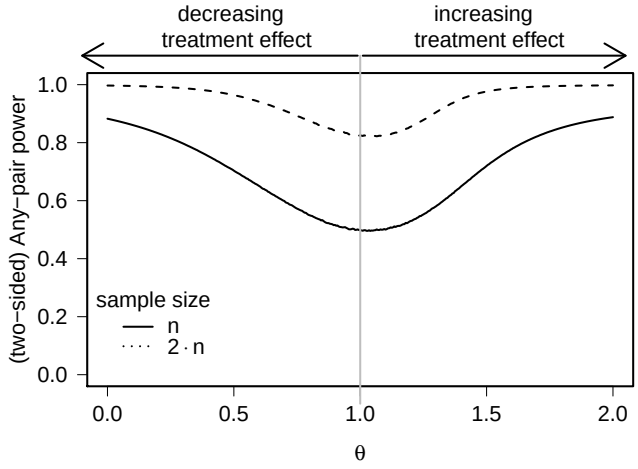

Figure 3. Power to detect at least one inconsistent treatment effect, where the treatment effects are defined as $\pi_{2j} - \pi_{1j}$ and $\pi_{3j} - \pi_{1j}$, among the three centres in the BLISS study depending on the consistency margin θ. Several sample sizes were considered: n - original sample size; 2n - twice as the original sample size; 4n - quadruple the original sample size. Two-sided hypothesis are considered: $H_A^{1j} : \delta_{1j}/\delta \neq \theta$, that detect an inconsistent treatment effect that is not equal to θ

4. Discussion

In this paper, power and sample size formulas applicable for design considerations when interest is in simultaneously assessing consistency of the treatment effect in a subset of patients are provided. The underlying methodology depends on the formulation of the treatment-by-subset interaction as ratio of treatment differences. This flexible approach allows the application of the presented method on a wide variety of trials, e.g. i) assessing consistency in MRCTs (Chen et al., 2010), ii) analysis of treatment-by-subgroup interaction in subgroup analysis (Gail and Simon, 1985), iii) detecting qualitative interactions in multi-centre clinical trials (Potthoff et al., 2001), or iv) non-inferiority analysis in multi-arm studies (Dilba et al., 2006).

In a two-arm clinical trial it is common practise to define a treatment effect as the difference between the two treatment levels (Gail and Simon, 1985; Quan et al., 2010a; Kitsche and Hothorn, 2014). In a multi-arm study the treatment effects are defined as user-specified linear combinations of the treatment levels. The linear combinations have to be chosen such that they reflect the research question, see e.g. Bretz and Hothorn (2002) for several examples in a one-way layout. The application on a three-arm trial is demonstrated by using the BLISS study example within this paper. Furthermore, it has to be noted that cases where the number of subsets reduces to $j = 1$ the presented methodology simplifies to a non-inferiority problem. In that particular case it coincides with the method of Hauschke and Kieser (2001) and Dilba et al. (2006) for continuous response variables and the method presented by Bretz and Hothorn (2002) for the binomial case.

Within this manuscript the consistency criterion of Method 1 from the Japanese MHLW and Definition 1 from Quan et al. (2010a) are used to examine whether the overall results from a MRCT can be applied to all regions. Among the many power definitions that can occur in multiple hypotheses testing, formulas for the any-pair and the all-pair power are presented. The any-pair power corresponds to the probability of detecting at least one inconsistent treatment effect under the alternative hypotheses, whereas the all-pair power corresponds to the detection of all inconsistent treatment effects. A critical step is to select the magnitude of the consistency margin θ. This can be determined by the regulatory agency in the specific region. The Japanese MHLW suggests that $\theta = 0.5$ to detect an inconsistent treatment effect between the Japanese region and all regions. Nevertheless, as notes by Chen et al. (2010) a value of $\theta = 0.5$ may be too conservative and even not practical if more than two regions are included in the analysis. Therefore, they recommend a smaller value of $\theta = 1/J$, where J denotes the number of pre-defined regions. Using the proposed methodology it is also possible to define region-specific consistency margins within a vector $\boldsymbol{\theta}$.

The proposed methodology is presented to detect an inconsistent treatment effect. Nevertheless, if a researcher is interested in the power that all treatment effects are greater than a pre-specified consistency margin the null and alternative hypothesis in Eq. 3 have to be inverted and the all-pairs power definition has to be applied. However, it should be noted that the total sample size required to detect a significant consistent treatment effect in all regions could potentially be many-fold increase as compared with the traditional methods.

The presented procedure to detect an inconsistent treatment effect is applicable for quantitative and binary response variables. Nevertheless, the primary endpoint in many clinical trials is a survival endpoint. Quan et al. (2010b) provide sample size formulas for survival outcome, but their method is limited to Japanese patients in MRCTs. Future research

could focus on the combination of the procedure for many-to-one comparisons in the frailty Cox model proposed by Herberich and Hothorn (2012) and the ratio formulation of the treatment-by-subset interaction presented by Kitsche and Hothorn (2014).

References

Agresti, A. and Ryu, E. (2010). Pseudo-score confidence intervals for parameters in discrete statistical models. *Biometrika*, 97(1):215–222.

Bretz, F., Genz, A., and Hothorn, L. A. (2001). On the Numerical Availability of Multiple Comparison Procedures. *Biometrical Journal*, 43(5):645–656.

Bretz, F. and Hothorn, L. A. (2002). Detecting dose-response using contrasts: asymptotic power and sample size determination for binomial data. *Statistics in Medicine*, 21(22):3325–35.

Chen, J., Quan, H., Binkowitz, B., Ouyang, S., 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., Zheng, H., Quan, H., Li, G., Gallo, P., Ouyang, S. P., Binkowitz, B., Ting, N., Tanaka, Y., Luo, X., and Ibia, E. (2013). Graphical assessment of consistency in treatment effect among countries in multi-regional clinical trials. *Clinical Trials*, 10(6):842–51.

Dilba, G., Bretz, F., Hothorn, L. A., and Guiard, V. (2006). Power and sample size computations in simultaneous tests for non-inferiority based on relative margins. *Statistics in Medicine*, 25(7):1131–47.

Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., and Offen, W. (2005). *Analysis ofClinical Trials using SAS: A Practical Guide*. Cary, NC, SAS Institute Inc.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41(2):361–72.

Genz, A. and Bretz, F. (1999). Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63(4):103–117.

Hauschke, D. and Kieser, M. (2001). Multiple testing to establish noninferiority of k treatments with a reference based on the ratio of two means. *Drug information journal*, 35:1247–1251.

Hauschke, D., Kieser, M., Diletti, E., and Burke, M. (1999). Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Statistics in Medicine*, 18(1):93–105.

Herberich, E. and Hothorn, T. (2012). Dunnett-type inference in the frailty Cox model with covariates. *Statistics in medicine*, 31(1):45–55.

Horn, M. and Vollandt, R. (1998). Sample Sizes for Comparisons ofk Treatments with a Control Based on Different Definitions of the Power. *Biometrical Journal*, 40(5):589–612.

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

Kitsche, A. (2014). Detecting qualitative interactions in clinical trials with binary responses. *Pharmaceutical statistics*.

Kitsche, A. and Hothorn, L. A. (2014). Testing for qualitative interaction using ratios of treatment differences. *Statistics in Medicine*, 33:1477–1489.

Li, M., Quan, H., Chen, J., Tanaka, Y., Ouyang, P., Luo, X., and Li, G. (2012). R Functions for Sample Size and Probability Calculations for Assessing Consistency of Treatment Effects in Multi-Regional Clinical Trials. *Journal of Statistical Software*, 47:1–10.

MERIT-HF Study Group (1999). Effect of metoprolol CR/XL in chronic heart failure: Metoprolol CR/XL Randomised Intervention Trial in-Congestive Heart Failure (MERIT-HF). *The Lancet*, 353(9169):2001–2007.

Ministry of Health Labour and Welfare of Japan (2007). Basic concepts for joint international clinical trials.

Navarra, S. V., Guzmán, R. M., Gallacher, A. E., Hall, S., Levy, R. A., Jimenez, R. E., Li, E. K.-M., Thomas, M., Kim, H.-Y., León, M. G., Tanasescu, C., Nasonov, E., Lan, J.-L., Pineda, L., Zhong, Z. J., Freimuth, W., and Petri, M. A. (2011). Efficacy and safety of belimumab in patients with active systemic lupus erythematosus: a randomised, placebo-controlled, phase 3 trial. *Lancet*, 377(9767):721–31.

Potthoff, R. F., Peterson, B. L., and George, S. L. (2001). Detecting treatment-by-centre interaction in multi-centre clinical trials. *Statistics in Medicine*, 20(2):193–213.

Quan, H., Li, M., Chen, J., Gallo, P., Binkowitz, B., Ibia, E., Tanaka, Y., Ouyang, S. P., Luo, X., Li, G., Menjoge, S., Talerico, S., and Ikeda, K. (2010a). Assessment of consistency of treatment effects in multiregional clinical trials. *Drug Information Journal*, 44:617–632.

Quan, H., Li, M., Shih, W. J., Ouyang, S. P., Chen, J., Zhang, J., and Zhao, P.-L. (2013). Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials. *Statistics in Medicine*, 32(10):1691–706.

Quan, H., Zhao, P., 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(2):100–112.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tsou, H.-H., James Hung, H. M., Chen, Y.-M., Huang, W.-S., Chang, W.-J., and Hsiao, C.-F. (2012). Establishing consistency across all regions in a multi-regional clinical trial. *Pharmaceutical Statistics*, 11(4):295–9.

Wittes, J. (2013). Why Is This Subgroup Different from All Other Subgroups? Thoughts on Regional Differences in Randomized Clinical Trials. In Fleming, T. R. and Weir, B. S., editors, *Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials*, volume 1205 of *Lecture Notes in Statistics*, pages 95–115, New York, NY. Springer New York.

A. Power and sample size for binary response data

B. General form of denominator interaction contrast matrix in Eq. 11

$$\boldsymbol{C}_{\text{Denominator}} = \begin{pmatrix} -\frac{n_{1,12}}{N_{12}} & \frac{n_{1,12}}{N_{12}} & 0 & -\frac{n_{2,12}}{N_{12}} & \frac{n_{2,12}}{N_{12}} & 0 & -\frac{n_{3,12}}{N_{12}} & \frac{n_{3,12}}{N_{12}} & 0 \\ -\frac{n_{1,13}}{N_{13}} & 0 & \frac{n_{1,13}}{N_{13}} & -\frac{n_{2,13}}{N_{13}} & 0 & \frac{n_{2,13}}{N_{13}} & -\frac{n_{3,13}}{N_{13}} & 0 & \frac{n_{3,13}}{N_{13}} \\ -\frac{n_{1,12}}{N_{12}} & \frac{n_{1,12}}{N_{12}} & 0 & -\frac{n_{2,12}}{N_{12}} & \frac{n_{2,12}}{N_{12}} & 0 & -\frac{n_{3,12}}{N_{12}} & \frac{n_{3,12}}{N_{12}} & 0 \\ -\frac{n_{1,13}}{N_{13}} & 0 & \frac{n_{1,13}}{N_{13}} & -\frac{n_{2,13}}{N_{13}} & 0 & \frac{n_{2,13}}{N_{13}} & -\frac{n_{3,13}}{N_{13}} & 0 & \frac{n_{3,13}}{N_{13}} \\ -\frac{n_{1,12}}{N_{12}} & \frac{n_{1,12}}{N_{12}} & 0 & -\frac{n_{2,12}}{N_{12}} & \frac{n_{2,12}}{N_{12}} & 0 & -\frac{n_{3,12}}{N_{12}} & \frac{n_{3,12}}{N_{12}} & 0 \\ -\frac{n_{1,13}}{N_{13}} & 0 & \frac{n_{1,13}}{N_{13}} & -\frac{n_{2,13}}{N_{13}} & 0 & \frac{n_{2,13}}{N_{13}} & -\frac{n_{3,13}}{N_{13}} & 0 & \frac{n_{3,13}}{N_{13}} \end{pmatrix}, \tag{12}$$

where $N_{jj'} = \sum_{j \in j,j'} \sum_{i=1}^{I} n_{ij}$ and $n_{i,jj'} = \sum_{j \in j,j'} n_{ij}$.