Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2014 *Paper* 108

Methods for Exploring Treatment Effect Heterogeneity in Subgroup Analysis: An Application to Global Clinical Trials

I. Manjula Schou* Ian C. Marschner†

*Macquarie University, IM.SCHOU@Yahoo.com.au

†Macquarie University, ian.marschner@mq.edu.au

Methods for Exploring Treatment Effect Heterogeneity in Subgroup Analysis: An Application to Global Clinical Trials

I. Manjula Schou and Ian C. Marschner

Abstract

Multi-country randomised clinical trials (MRCTs) are common in the medical literature and their interpretation has been the subject of extensive recent discussion. In many MRCTs, an evaluation of treatment effect homogeneity across countries or regions is conducted. Subgroup analysis principles require a significant test of interaction in order to claim heterogeneity of treatment effect across subgroups, such as countries in a MRCT. As clinical trials are typically underpowered for tests of interaction, overly optimistic expectations of treatment effect homogeneity can lead researchers, regulators and other stakeholders to over-interpret apparent differences between subgroups even when heterogeneity tests are insignificant. In this paper we consider some exploratory analysis tools to address this issue. We present three measures derived using the theory of order statistics which can be used to understand the magnitude and the nature of the variation in treatment effects that can arise merely as an artefact of chance. These measures are not intended to replace a formal test of interaction, but instead provide non-inferential visual aids allowing comparison of the observed and expected differences between regions or other subgroups, and are a useful supplement to a formal test of interaction. We discuss how our methodology differs from recently published methods addressing the same issue. A case study of our approach is presented using data from the PLATO study, which was a large cardiovascular MRCT that has been the subject of controversy in the literature. An R package is available from the authors on request.

Methods for exploring treatment effect heterogeneity in subgroup analysis: an application to global clinical trials

I. Manjula Schou[a,b] and Ian C. Marschner[a,b,*]

[a] Department of Statistics, Macquarie University, NSW 2109, Australia.
[b] NHMRC Clinical Trials Centre, University of Sydney, NSW 2006, Australia.
[*] Corresponding author: Ian C. Marschner, Department of Statistics, Macquarie University, NSW 2109, Australia. *E-mail: ian.marschner@mq.edu.au*

Abstract

Multi-country randomised clinical trials (MRCTs) are common in the medical literature and their interpretation has been the subject of extensive recent discussion. In many MRCTs, an evaluation of treatment effect homogeneity across countries or regions is conducted. Subgroup analysis principles require a significant test of interaction in order to claim heterogeneity of treatment effect across subgroups, such as countries in a MRCT. As clinical trials are typically underpowered for tests of interaction, overly optimistic expectations of treatment effect homogeneity can lead researchers, regulators and other stakeholders to over-interpret apparent differences between subgroups even when heterogeneity tests are insignificant. In this paper we consider some exploratory analysis tools to address this issue. We present three measures derived using the theory of order statistics which can be used to understand the magnitude and the nature of the variation in treatment effects that can arise merely as an artefact of chance. These measures are not intended to replace a formal test of interaction, but instead provide non-inferential visual aids allowing comparison of the observed and expected differences between regions or other subgroups, and are a useful supplement to a formal test of interaction. We discuss how our methodology differs from recently published methods addressing the same issue. A case study of our approach is presented using data from the PLATO study, which was a large cardiovascular MRCT that has been the subject of controversy in the literature. An R package is available from the authors on request.

Keywords: clinical trial; heterogeneity; interaction; multi-country study; subgroup analysis

1 Introduction

Multi-country randomised clinical trials (MRCTs) evaluating new drug therapies are popular as they efficiently pool resources to provide faster recruitment and more generalisable results across patient populations, ethnicities and disease management

paradigms. MRCTs also have the advantage of providing country-specific data that can be used for local regulatory dossiers that may otherwise require bridging studies, and local reimbursement applications for countries that have government funded pharmaceutical schemes. As a supplement to the overall analysis, MRCTs often present country-specific results that effectively correspond to a subgroup analysis. These subgroups may be defined by the individual countries participating in the study, or by pooling several countries in a geographical region, to avoid issues of low power or analytical complications that can arise from low enrolment in individual countries. In this paper we will focus on the interpretation of these country- or region-specific subgroup analyses, and will use the terms country and region interchangeably.

Clinical trials often assess the consistency of the treatment effect across pre-specified subgroups and generally accepted principles of subgroup analysis have been developed [1, 2]. In MRCTs, there is typically an assessment of treatment effect homogeneity across subgroups defined by regions. According to subgroup analysis principles, a test of interaction is the standard assessment of treatment effect heterogeneity across subgroups. However, as most studies are only designed with adequate power to detect an overall clinically meaningful difference between treatments in the primary endpoint, the test of interaction to assess heterogeneity of treatment effect across regions in a MRCT can often be underpowered [3, 4]. Indeed, the power decreases further as the number of regions in the subgroup analysis increases. Therefore, when there is a non-significant p-value from a test of interaction in the presence of seemingly heterogeneous treatment effects across regions, speculation of a type II error can arise making interpretation of the regional results difficult. It is very important that such speculation takes due account of the fact that random variation can result in some regions showing a lack of benefit even when there is no underlying heterogeneity and the treatment effect is beneficial. To this end, it is prudent to investigate whether potential differences exist between regions that can plausibly lead to differential treatment benefit and to appropriately design a study with this in mind [5]. A design paper or the study analysis plan can also be used to pre-emptively document and help inform researchers of the extent of chance variation to anticipate in a planned MRCT [6].

Consideration of potential differences between region-specific treatment effects is important at both the design stage and the analysis stage of a MRCT. At the design stage it is useful to understand the nature and extent of chance differences that can be expected to arise between regions, under the assumption of treatment effect homogeneity. At the analysis stage it is useful to compare the observed regional treatment differences with the expected regional treatment differences and assess the magnitude of any differences. As such, the intent of this approach is not to determine how the methodology performs under heterogeneity. Instead, it assesses the potential extent of chance variation under an assumption of homogeneity. A previous paper focused on considerations at the design stage [6], and in the present paper we adapt and extend this approach for application at the analysis stage. Our methods are based on the theory of order statistics for heteroscedastic normally distributed variables, which is applied to the collection of region-specific treatment differences. This allows various comparisons of expected subgroup-specific effects to be made with the actual observed effects under an assumption of treatment effect homogeneity. Specifically, we investigate the expected and observed effects via a comparison of order statistics, the probability of subgroups favouring the control, and the distribution of the range of treatment effects.

The resulting collection of graphical presentations provides a useful supplementary tool to the test of interaction and can equip researchers with a visual summary of the concordance between the observed treatment differences across regions and those expected due to chance. Although we will focus on regional differences in MRCTs, the methodology that we propose is equally applicable to other subgroup analyses.

Over recent years there has been a high level of research activity on statistical considerations relating to treatment effect heterogeneity in MRCTs and multi-centre studies, reflecting the practical importance of this issue [4, 5, 6, 7, 8, 9, 11, 12, 13]. In the next section we will begin by reviewing past methods of relevance to those discussed here, including a very recently published approach which, like ours, is based on the theory of order statistics [13]. We then introduce our methodological extensions, as well as providing a discussion of the fundamental differences between our approach and previous approaches, particularly our use of absolute treatment effects rather than standardised treatment effects in assessing the concordance between observed and expected treatment effect heterogeneity. Although this introduces methodological complexities compared to past approaches [13], we argue that this leads to more interpretable exploratory analysis tools. Finally we consider a detailed case study of the methods based on the PLATO study, which was a large MRCT of ticagrelor and clopidogrel for the prevention of cardiovascular events in patients with acute coronary syndromes [14]. Application of our methods to the PLATO study, which has been the subject of much discussion in the literature, suggests that the apparently large variation in country-specific treatment effects is consistent with the play of chance.

2 Overview of previous research

We begin with an overview of previous work which our research extends, together with an introduction of the assumptions and notation that will be used throughout the paper.

2.1 Assumptions

Consider the comparison of two treatment groups, a control treatment group and an experimental treatment group, in a MRCT conducted over R regions. The sample size for treatment group i in region r is n_{ir}, for $i = 1, 2$ and $r = 1, \ldots, R$. It is assumed that there is a parameter δ which measures the treatment effect, with $\delta = 0$ corresponding to no difference between the treatments. In principle the parameter δ could depend on r, meaning that there is genuine treatment effect heterogeneity across the regions. However, here we will make the assumption that δ does not depend on r, because our methods are aimed at assessing the extent of chance variation that could arise in the observed region-specific treatment effects under the assumption that there is underlying homogeneity.

The treatment effect δ could take a variety of forms depending on the type of primary endpoint that is being used in the MRCT. For example, with continuous endpoints δ may be a mean difference, with binary endpoints δ may be a risk difference, log relative risk or log odds ratio, while for time-to-event endpoints δ may be a log hazard ratio. Regardless of the type of treatment effect that δ measures, it will be assumed that for each region there is a region-specific estimator D_r of δ, which has a

normal distribution

$$D_r \sim N\left(\delta, s_r^2\right) \qquad r = 1, \ldots, R. \tag{1}$$

This distributional assumption will be reasonable for most types of treatment effect measures on an appropriate scale, at least under a large sample assumption with approximate normality. Furthermore, it is assumed that the region-specific estimators are independent random variables. Other than these general assumptions it is not necessary for us to make any specific assumptions about the type of endpoint or the treatment effect measure δ. In the case study described in Section 4 we will make use of the above model with a time-to-event endpoint where δ is a log hazard ratio parameter and D_r are country-specific log-hazard ratio estimators. However, it is also applicable for other treatment effect measures, and has been used for relative risks and risk differences in other contexts [6, 15].

The form of s_r^2 in (1) can be derived in terms of the proportion of the study enrolment allocated to region r and the design parameters used in the overall sample size calculation, including the power, significance level and the homogeneous treatment difference δ. This form of s_r is useful for the assessment of expected treatment effect heterogeneity at the design stage, as illustrated by Marschner [6]. At the analysis stage s_r^2 will not be known in general, so a standard error estimate must also be available as discussed further in Section 3.5.

2.2 Expected range

Marschner [6] proposed the expected range of region-specific treatment effects as a useful benchmark for the expected treatment effect variation. The expected range can be derived based on the distribution function of the smallest and largest order statistics, $D_{(1)}$ and $D_{(R)}$, which are respectively

$$F_{(1)}(x) = 1 - \prod_{i=1}^{R} \{1 - F_i(x)\} = 1 - \prod_{i=1}^{R} \left\{1 - \Phi\left(\frac{x-\delta}{s_i}\right)\right\}$$

and

$$F_{(R)}(x) = \prod_{i=1}^{R} F_i(x) = \prod_{i=1}^{R} \Phi\left(\frac{x-\delta}{s_i}\right).$$

Here, F_i is the distribution function of the normal distribution in equation (1) with $r = i$, while Φ is standard normal distribution function.

Using these distribution functions, the expectations of $D_{(1)}$ and $D_{(R)}$ can be calculated, as can the expectation of the range of treatment effects, $V = D_{(R)} - D_{(1)}$ [6]. The expectation $E(V)$ provides a measure of the range of the treatment differences that can be expected due to chance, under an assumption of treatment effect homogeneity across the regions. The intent of this measure was to facilitate a comparison of the range of observed and expected treatment differences, thus providing a non-inferential complement to the primary assessment based on a test of interaction of treatment effect differences across regions. Subsequently the expected range has also been used in a more inferential capacity by Chen et al. [13], although this was not the original intention.

2.3 Probability of at least one region favouring the control

An alternative measure that is also based on the extreme order statistics and provides information about the expected variation in region-specific treatment effects is the probability of at least one region favouring the control [6, 16]. The motivation for considering this quantity is that an inconsistent region-specific treatment effect in a study that shows an overall benefit in favour of the experimental treatment will often prompt further investigation and interpretation. Quantifying the probability of this event, and the extent to which it is likely or unlikely, therefore provides a benchmark against which the occurrence of an inconsistent region-specific treatment effect can be interpreted.

Assuming δ is scaled such that a negative value for the treatment difference indicates benefit in favour of the experimental treatment, then the probability of at least one region favouring the control is given by

$$\Pr\left(D_{(R)} > 0\right) = 1 - \prod_{i=1}^{R} F_i(0) = 1 - \prod_{i=1}^{R} \Phi\left(\frac{-\delta}{s_i}\right).$$

As with the expected range, the intent of this measure is to provide a non-inferential tool to calibrate expectations about whether all treatment effects should lie in a consistent direction. If the probability is substantial, then it should not be too surprising if an inconsistent treatment effect is observed in a particular region, and over-interpretation of such an observation should be avoided. Such information can be taken into consideration alongside the test of interaction.

2.4 Normal scores

While the extreme order statistics $D_{(1)}$ and $D_{(R)}$ provide information about treatment effect heterogeneity, it is natural to consider more informative methods based on all order statistics. A recently proposed alternative approach of Chen et al. [13] does this. This approach assesses treatment effect heterogeneity using normal probability plots comparing the ordered standardised treatment differences with their associated normal scores. Specifically, the approach uses the standardised quantity referred to as the weighted least squares residual defined as $e_r = (D_r - \hat{\delta})/s_r$. Here

$$\hat{\delta} = \sum_{r=1}^{R} w_r D_r$$

is an unbiased estimator of δ with the weights $w_r = s_r^{-2} / \left(\sum_{i=1}^{R} s_i^{-2}\right)$ reflecting the amount of statistical information provided by region r, or equivalently the precision of the region-specific estimator D_r. Under the assumption of treatment effect homogeneity, the weighted least squares residuals are distributed as

$$e_r = \frac{(D_r - \hat{\delta})}{s_r} \sim N\big(0, 1 - w_r\big). \tag{2}$$

It then follows from (2) that the standardised weighted least squares residual $\widetilde{e}_r = e_r/\sqrt{1 - w_r}$ has a standard normal distribution. The method proposes comparing

the ordered standardised weighted least squares residuals $\widetilde{e}_{(r)}$, $r = 1, \ldots, R$, with the standard normal scores which can be readily obtained using standard tables or software [17, 18]. The main tool for undertaking this comparison is a normal probability plot. In the special case of a homoscedastic normal outcome where δ is the mean difference and the treatment group sizes are equal within each region, the weights w_r reduce to the proportion of the overall sample size that comes from region r [13]. However, the above approach applies more generally, and can be used for other treatment difference measures that conform with the basic assumption (1).

In the present paper, our most significant contribution is to adapt this normal scores method to make use of the absolute order statistics $D_{(r)}$ in place of the standardised order statistics $\widetilde{e}_{(r)}$. In the next section we consider the substantial methodological complexities this introduces, but also explain why we believe this leads to a more interpretable assessment of treatment effect heterogeneity.

3 Methodological extensions

In this section we consider various extensions and adaptations of the methods reviewed in the previous section. We will focus on three measures that can be used in comparing the observed variation in treatment effects with what would be expected by chance under the assumption of treatment effect homogeneity across regions.

3.1 Overview of extensions

The first of the three measures we consider is the expected value of the r^{th} order statistic of the region-specific treatment effects, $E\big(D_{(r)}\big)$, for each $r = 1, \ldots, R$. Comparison of these expected order statistics with the sample order statistics $D_{(r)}$, for example using a normal probability plot, provides an alternative version of the comparison described in Section 2.4, between $\widetilde{e}_{(r)}$ and the normal scores. Although it may seem like a natural alternative to use of the absolute treatment effects rather than the standardised treatment effects, this introduces a number of complexities because the $D_{(r)}$ quantities are the order statistics from a heteroscedastic sample. These complexities are addressed in the next section. Despite the additional complexity we argue in Section 3.4 that this comparison provides a preferable assessment of treatment effect heterogeneity than the use of standardised treatment effects as used by Chen et al. [13].

The second measure involves using the full distribution of the number of regions that favour the control, rather than the more restrictive quantity discussed in Section 2.3, the probability of at least one region favouring the control. This distribution will be helpful in interpreting studies where more than one region favours the control, which is not uncommon in MRCTs involving a large number of regions.

Finally, the third measure we consider is the full probability distribution of the treatment effect range, $D_{(R)} - D_{(1)}$, which is helpful in interpreting the treatment effect range observed in a MRCT. Use of the full distribution generalises the expected range approach described in Section 2.2, which is based just on the expected value of the effect range distribution. In principle this approach could also be generalised to other range-based distributions, such as the distribution of the inter-quartile range of region-specific treatment effects. Here, however, we restrict our focus to the range

of treatment effects which, as described in Section 2.2, has been the focus of prior research.

3.2 Order statistic distribution

All of our methods depend fundamentally on the distribution of the order statistics of the region-specific treatment effects. This involves considering the order statistics from a sample of R heteroscedastic normal variates. We now consider this distribution and then describe how it can be used to derive the three measures of expected treatment effect heterogeneity.

The distribution function of $D_{(r)}$ is

$$\begin{aligned} F_{(r)}(x) &= \Pr\left(D_{(r)} \leq x\right) \\ &= \Pr\left(\text{At least } r \text{ of } R \text{ treatment differences do not exceed } x\right) \\ &= \sum_{i=r}^{R} \sum_{S \in S_i(R)} \left\{ \prod_{k \in S} F_k(x) \prod_{\substack{k=1 \\ k \notin S}}^{R} \left[1 - F_k(x)\right] \right\} \end{aligned} \tag{3}$$

where $S_i(R)$ is the family of all subsets of size i from $\{1, \ldots, R\}$ [19].

On expansion and simplification of (3) we get

$$F_{(r)}(x) = \sum_{i=r}^{R} c_{ir} \sum_{S \in S_i(R)} \prod_{k \in S} F_k(x) \tag{4}$$

where

$$c_{ir} = (-1)^{i-r} \binom{i-1}{r-1}.$$

In the special case where the D_rs are independent identically distributed random variables with $s_r = s$, equation (4) reduces to

$$F_{(r)}(x) = \sum_{i=r}^{R} c_{ir} \binom{R}{i} F(x)^i,$$

and is equivalent to the familiar representation [17]

$$F_{(r)}(x) = \sum_{i=r}^{R} \binom{R}{i} F(x)^i \left[1 - F(x)\right]^{R-i}.$$

However, our formulation allows for a fully heteroscedastic specification which is required to allow for different regions having different sample sizes.

Applying the product rule for differentiation on the distribution function, the probability density of the r^{th} order statistic is

$$f_{(r)}(x) = \sum_{i=r}^{R} \sum_{j=1}^{R} c_{ir} f_j(x) \sum_{S \in S_i(R)} 1\{j \in S_i(R)\} \prod_{\substack{k \in S \\ k \neq j}} F_k(x) \tag{5}$$

where $1\{\cdot\}$ is the indicator function. Although this theoretical specification appears unwieldy, it is straightforward to compute.

As with $F_{(r)}(x)$, a simplified version of (5) is achieved in the special case where the D_rs are independent identically distributed random variables with $s_r = s$, and is given by

$$f_{(r)}(x) = \sum_{i=r}^{R} i c_{ir} \binom{R}{i} f(x) F(x)^{i-1}.$$

A simplified illustrative example of the order statistic distribution is provided for a MRCT with $R = 3$ regions and treatment differences D_1, D_2, and D_3. In this case, consider the distribution of $D_{(2)}$. Here, the family of sets $S_2(3)$ and $S_3(3)$ would be given by $S_2(3) = \{\{1,2\},\{1,3\},\{2,3\}\}$ and $S_3(3) = \{\{1,2,3\}\}$. The distribution and density functions of $D_{(2)}$ follow readily from (4) and (5) and the fact that $c_{22} = 1$ and $c_{32} = -2$, namely

$$\begin{aligned} F_{(2)}(x) &= F_1(x)F_2(x)\left[1 - F_3(x)\right] + F_1(x)F_3(x)\left[1 - F_2(x)\right] \\ &+ F_2(x)F_3(x)\left[1 - F_1(x)\right] + F_1(x)F_2(x)F_3(x) \\ &= c_{22} \times \{F_1(x)F_2(x) + F_1(x)F_3(x) + F_2(x)F_3(x)\} \\ &+ c_{32} \times \{F_1(x)F_2(x)F_3(x)\} \end{aligned}$$

and

$$\begin{aligned} f_{(2)}(x) &= c_{22} \times \{f_1(x)F_2(x) + f_1(x)F_3(x) + f_2(x)F_1(x) \\ &\quad + f_2(x)F_3(x) + f_3(x)F_1(x) + f_3(x)F_2(x)\} \\ &+ c_{32} \times \{f_1(x)F_2(x)F_3(x) + f_2(x)F_1(x)F_3(x) + f_3(x)F_1(x)F_2(x)\}. \end{aligned}$$

The forms of $F_{(2)}(x)$ and $f_{(2)}(x)$ in this simplified 3-region example illustrate the link between equations (3) and (4) and the role of the c_{ir} constants in specifying the order statistic distribution.

As foreshadowed in Section 3.1, the general order statistic distribution for heteroscedastic treatment effects can now be used to derive several useful measures of chance treatment effect variation that extend and improve upon the measures discussed in Section 2.

3.3 Measures of chance variation

The first measure described in Section 3.1, the expectation of the r^{th} order statistic of the region-specific treatment differences, can now be obtained using (5)

$$E\big(D_{(r)}\big) = \int_{-\infty}^{\infty} x f_{(r)}(x) dx. \tag{6}$$

Although this form is not explicit, it can be straightforwardly computed using standard routines for numerical integration. As explained later in the paper, all computations presented here were performed in R [18].

Once computed, these expected order statistics can be compared graphically with the observed ordered treatment differences to assess whether the observed spread of region-specific treatment differences is unusual relative to what would be expected by chance under the assumption of treatment effect homogeneity. One such plot would a simple box plot of the observed and expected order statistics which provides a graphical

generalisation of the approach of comparing the observed and expected ranges [6]. Another approach would be a plot of the observed versus expected treatment differences, which is a type of normal probability plot that conveys information about any departures of the observed region-specific treatment effects from what would be expected by chance.

Treatment differences that align consistently across regions in terms of direction of effect are straightforward to interpret and explain. Sometimes, however, chance variation will lead some regions to have a treatment effect estimate that goes in the opposite direction to the overall effect. This can potentially lead to speculation and over-interpretation. Therefore, being able to compare the observed number of regions favouring the control with the probability distribution of the number of regions favouring the control provides a useful benchmark by which to assess the role of chance variation. This leads to the second of the three approaches introduced in Section 3.1, which generalises the previously suggested approach discussed in Section 2.3.

Like the other quantities discussed in this section, the probability distribution of W, the number of regions favouring the control, is connected to the order statistic distribution discussed in Section 3.2 through the relationship

$$\Pr(W \geq w) = \Pr\big(D_{(R-w+1)} > 0\big) = 1 - F_{(R-w+1)}(0) \qquad w = 1, \ldots, R. \tag{7}$$

Assuming a positive treatment difference signifies an effect in favour of the control treatment, and letting p_i be the probability that region i favours the control, we have the following

$$p_i = \Pr\big(D_i > 0\big) = 1 - F_i(0) = 1 - \Phi\left(-\delta/s_i\right).$$

It then follows from equations (3) and (7) that the probability function of the number of regions favouring the control is

$$\begin{aligned} P_W(w) &= \Pr(W = w) = \Pr(W \geq w) - \Pr(W \geq w+1) \\ &= F_{(R-w)}(0) - F_{(R-w+1)}(0) \\ &= \sum_{S \in S_w(R)} \prod_{k \in S} p_k \prod_{\substack{l=1 \\ l \notin S}}^{R} (1 - p_l) \qquad w = 0, \ldots, R \end{aligned} \tag{8}$$

where, for notational purposes, we define $F_{(0)}(0) = 1$ and $F_{(R+1)}(0) = 0$. For example, in the 3-region illustration discussed previously, the probability that two regions favour the control is

$$P_W(2) = \Pr(W = 2) = p_1p_2(1 - p_3) + p_1p_3(1 - p_2) + p_2p_3(1 - p_1).$$

Once this distribution has been computed, the observed number of regions favouring the control can be compared with $P_W(w)$ in order to assess whether the observed number is unusual compared with what would be expected by chance under the assumption of homogeneous treatment effects. A natural summary measure of the extent to which the observation $W = w_o$ is consistent with chance variation, is the probability of obtaining an observation at least as extreme as $W = w_o$, namely, $P_E = \Pr\big(W \geq w_o\big)$. Although we are not recommending P_E as a p-value for formal hypothesis testing, it does nonetheless provide a non-inferential quantification of the extent to which the observed number of inconsistent regions is unusual relative to what would be expected by chance.

Finally, the third approach introduced in Section 3.1 is based on the probability distribution of the range of region-specific treatment effects, $V = D_{(R)} - D_{(1)}$. This distribution is well known in the homoscedastic case based on the joint distribution of $D_{(1)}$ and $D_{(R)}$ [17]. In the heteroscedastic generalisation that we are using in this paper, the density function of the range can be expressed as follows for $x \geq 0$.

$$f_V(x) = \int_{-\infty}^{\infty} \sum_{i=1}^{R} \sum_{\substack{j=1 \\ j \neq i}}^{R} f_i(y) f_j(y+x) \prod_{\substack{k=1 \\ k \neq i,j}}^{R} [F_k(y+x) - F_k(y)] \, dy. \quad (9)$$

As in equation (6), equation (9) requires numerical integration which we have undertaken in R using the `integrate` routine [18]. Once computed, the observed range can be compared with the probability distribution $f_V(x)$ to assess whether the observed range of treatment effects is unusual relative to what would be expected by chance under an assumption of treatment effect homogeneity. As with W above, a natural summary measure of the extent to which the observation $V = v_o$ is consistent with chance variation, is provided by the probability of obtaining an observation at least as extreme as $V = v_o$, which in this case is $P_E = \int_{v_o}^{\infty} f_V(x)dx$.

An R package that implements all three approaches can be made available by the authors on request.

3.4 Comparison of the methods

While our methods and those of Chen et al. [13] both make use of assessments that are based on the theory of order statistics, there are important differences between the two approaches. Most significantly, our approach uses the observed region-specific treatment differences whereas the approach proposed by Chen et al. [13] uses the standardised treatment differences in the form of the weighted least squares residuals. In view of these differences, a discussion of the distinction between the two approaches is necessitated.

Statistically, the key distinction between using the absolute order statistics $D_{(r)}$ and the standardised order statistics $\widetilde{e}_{(r)}$, is that the former depends only on the treatment effects themselves, while the latter depends on a combination of the departure of the treatment effects from the overall effect and the associated standard error. Therefore, an ordering of the standardised weighted least squares residuals is essentially an ordering of the departure of the treatment effects from the overall effect, relative to the region-specific standard error, with the size of the standard error playing a critical role in the ordering. This may mean that the $D_{(r)}$ and $\widetilde{e}_{(r)}$ values are ordered in different ways. Indeed, this may mean that the two versions of order statistics convey different messages about whether the observed region-specific treatment effects are consistent with what would be expected by chance, and we provide an example of this in the case study discussed in Section 4.

The fact that the absolute and standardised treatment effects can convey different messages makes it important to consider how subgroup analyses are interpreted and used in practice by stakeholders. While the standardised treatment effects are what drives the formal test of heterogeneity, they are not the primary focus of subsequent informal assessments of the region-specific differences in treatment effects. Such informal assessments, which would typically follow an insignificant test of heterogeneity,

tend to focus on the absolute magnitudes of the treatment difference in each region. The spread in these absolute treatment effects is what then has the potential to lead to over-interpretation of apparent treatment effect variation. It therefore makes sense to focus on the expected variation in absolute treatment effects as a benchmark for the observed variation in absolute treatment effects. It is this use of actual rather than standardised treatment effects in the assessment and interpretation of heterogeneity that has led us to base our measures of expected variation on the actual treatment effects.

3.5 Implementation issues

In practice, there are several implementation issues that we discuss prior to considering a case study. Firstly, it requires noting that the various quantities discussed in the previous section are dependent on the unknown values of δ and s_r. This means that at the analysis stage of a study, sample estimates $\hat{\delta}$ and $\hat{s}_i$ are required so that computations of the expected variation in treatment effects can be undertaken. If the individual patient data are available, the overall treatment effect estimated using these data would be the most appropriate estimate of δ, as an assumption of treatment effect homogeneity underpins the assessment of chance variation. However, if only region-specific treatment effect estimates are available, the aggregated estimate of δ, as discussed in Section 2.4, would be used.

With regards to the standard error s_i, there are two possible approaches to estimation. The first, as used in this paper, would be to use the standard errors of the region-specific treatment effects as estimated separately within each region. This provides a more empirical estimate of standard error than the second approach which is to use the overall estimate of standard error, weighted by the proportion of subjects from each region. The latter approach enforces an assumption of region-level homoscedasticity and results in smaller regions being weighted less and larger regions being weighted more. This is a more natural approach to take at the design stage when no data is available.

A further implementation issue relates to the computational complexity of the methods. In Section 3.3 we presented theoretical expressions associated with the various measures of heterogeneity, that can be computed exactly with the aid of a routine to undertake numerical integration. In practice, it is also possible to approximate all of the required quantities using simulation. Although this is potentially computationally expensive, the computations themselves are trivial and obvious with the availability of a large number of simulated samples $D_1, \ldots, D_R$ from the normal distributions $N\big(\hat{\delta}, \hat{s}_r^2\big)$, for $r = 1, \ldots, R$. Since the theoretical computations required to compute the quantities described in Section 3.3 are based on combinatorial sets, there will generally be a point at which simulation becomes more efficient than direct computation. Based on our experience with the case study described in Section 4, the simulation approach tends to be preferable for $R > 20$.

4 Case study

4.1 PLATO study

As a case study, we consider the PLATO study which was a 43-country, double-blind, randomised trial comparing the experimental treatment ticagrelor with the control treatment clopidogrel, for the prevention of cardiovascular events in 18,624 subjects with acute coronary syndrome [14]. The primary endpoint of this study was the time to first occurrence of a cardiovascular event (death from vascular causes, myocardial infarction, or stroke). The study was designed to have 90% power to detect a relative risk reduction of 13.5%.

On completion, the overall study showed a significant reduction in cardiovascular events in favour of ticagrelor (hazard ratio 0.84, $p < 0.001$). Treatment effect heterogeneity was assessed in 33 separate subgroup analyses, one of which was an assessment of the heterogeneity of treatment effects across regions (Asia/Australia, Central/South America, Europe/Middle East/Africa and North America). The p-value for this test of interaction was 0.045 with the treatment effect in North America having an observed value that favoured the control, although insignificantly so. The investigators concluded that this finding may have been a chance result due to multiple testing, and that although no apparent explanations had been found, questioned whether the differences between patient populations and treatment practice patterns may have contributed to this result.

Although a p-value of 0.045 in the context of 33 subgroup analyses is not particularly surprising, the PLATO study was subsequently subjected to extensive post hoc analysis of country-specific heterogeneity in treatment effects. These analyses focused particularly on the observation that the USA treatment effect was in the direction favouring the control. The Food and Drug Administration (FDA) conducted its own review of the data following the sponsor's proposal of a potentially negative association between the dose of aspirin and the benefit of treatment with ticagrelor, finding that the dose of aspirin was higher in the USA subgroup compared with the non-USA subgroup [20]. A further review of this possible explanation was subsequently published together with a claim that differences in primary site monitoring by an independent contract research organisation (in the USA) and the study sponsor (in most other countries) may offer an alternative explanation requiring further investigation [21]. These proposals of a potential biological explanation (aspirin dose) and an operational explanation (site monitoring) were followed by a statistical assessment concluding that the country-specific treatment effect variation was consistent with the play of chance [22] and a further analysis concluding that the findings in the USA were likely not due to chance [13]. Here we use our methods to provide further exploration of the play of chance as a potential explanation for country-specific treatment effect differences in the PLATO study.

4.2 Data and analyses

In our analyses, we used published country-specific hazard ratios and 95% confidence intervals for all countries except the smallest (Hong Kong), which had only 16 patients. This led to $R = 42$ countries with sample sizes varying from 51 to 2666. We refer the reader to Figure 1 of Serebruany [21] for a full listing of the countries, sample sizes and

hazard ratios used in our analyses. The overall treatment effect δ was taken to be the log hazard ratio, for which assumption (1) is reasonable. The overall estimate $\hat{\delta}$ was calculated using an inverse variance weighting method based on country-specific log hazard ratios and standard errors calculated from the published confidence intervals.

As well as analyses of the treatment effects for all 42 countries, we also considered analyses restricted just to the countries with the largest sample sizes. These additional analyses served two purposes. Firstly, they enabled an assessment of the extent to which any conclusions are robust to the larger variation expected in small countries, which was raised as a concern by Chen et al. [13]. Secondly, these analyses served to illustrate the behaviour of the methodology on data sets having various R values. In our analyses we consider the results restricted to the largest 10, 15 and 20 countries, in addition to the full collection of 42 countries.

4.3 Order statistics

Figures 1 and 2 present the expected order statistics of the country-specific treatment differences displayed as box plots and normal probability plots. These plots are displayed for the entire collection of 42 countries, as well as analyses restricted to the largest 10, 15 or 20 countries. Also shown, in Figure 2 Panels B and D, are normal probability plots corresponding to the standardised weighted least squares residuals of Chen et al. [13], as discussed in Section 2.4. Since formal tests of heterogeneity of treatment effects are statistically insignificant ($p > 0.1$ in all cases), we intend that these graphical displays are used as a non-inferential supplement to a formal test of heterogeneity, in which the observed variation in treatment effects is compared with the expected variation in treatment effects. With this in mind, these figures do not identify any remarkable differences between what was observed and what would be expected due to chance variation. Figure 1 clearly displays the expected increase in treatment effect variation as more countries are included in the analysis, but does not suggest that the observed variation is inconsistent with what was expected under the hypothesis of homogeneity. Indeed, for the analyses involving larger numbers of countries (Panels C and D) it appears that the PLATO study actually exhibits less variation in country-specific treatment effects than would have been expected due to chance. This is also evident in Figure 2 Panel C, where the shallow gradient for all but the most extreme order statistics is indicative of smaller variation than expected.

Of particular interest is the comparison of Panels A and B of Figure 2, which is a comparison of the normal probability plots for absolute treatment effects and standardised treatment effects, for the analysis restricted to the largest 15 countries. Panel A, based on absolute treatment effects, displays no departure from the expected variation of treatment effects, with the possible exception of the smallest order statistics that suggest lower variation than expected. On the other hand, the standardised treatment effects displayed in Panel B show one outlying country, the USA, which seems to have a standardised treatment effect that departs from the other countries. This illustrates the potential for different qualitative messages to emerge from these two methods.

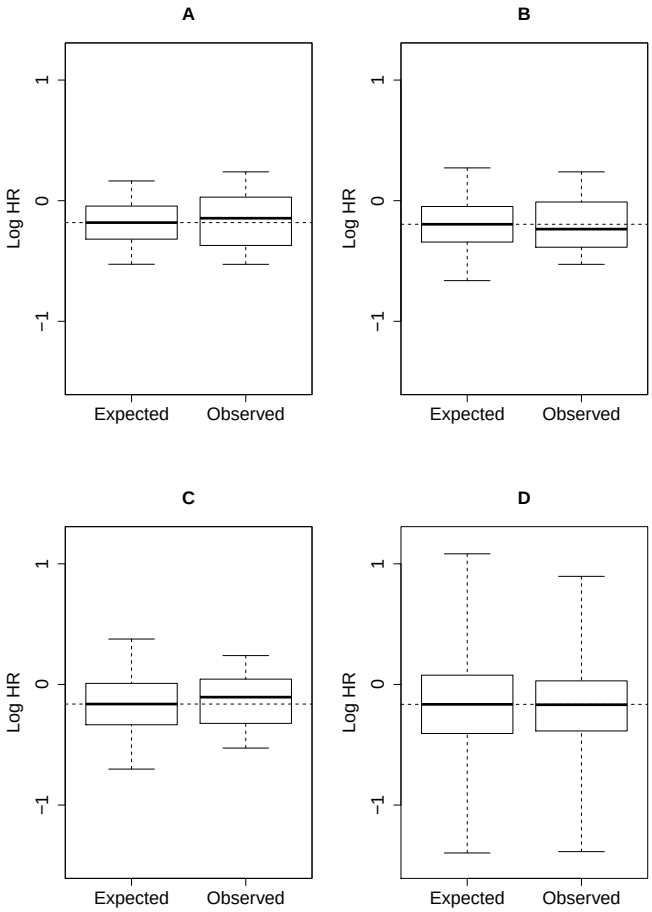

Figure 1: Observed and expected country-specific treatment differences from the PLATO study. The expected treatment differences for the largest 10 (Panel A), 15 (Panel B), 20 (Panel C) and 42 (Panel D) countries are plotted. The dotted line denotes the overall observed treatment difference.

4.4 Range of treatment effects

The expected range of treatment effects depicted in the extremities of the boxplots in Figure 1 can be generalised to the full distribution of the range of treatment effects, as discussed in Section 3.3. Plots of this distribution, together with the observed range of treatment effects, are provided in Figure 3. It can be seen that the observed range of country-specific treatment effects in the PLATO study is highly consistent with the distribution of the range of treatment effects under the assumption of treatment effect homogeneity. This conclusion is true regardless of whether analyses are restricted to the largest countries or include all countries. A useful summary measure of the extent of consistency is P_E, which was described in Section 3.3. In the present context, P_E is the probability of observing a treatment effect range at least as extreme as the one observed, under the assumption of treatment effect homogeneity. With $P_E = 0.55$, the overall analysis in Panel D of Figure 3 shows that a treatment effect range as large as the one observed in the PLATO study is highly likely, and could therefore plausibly have arisen through chance variation. The same conclusion would also be reached using

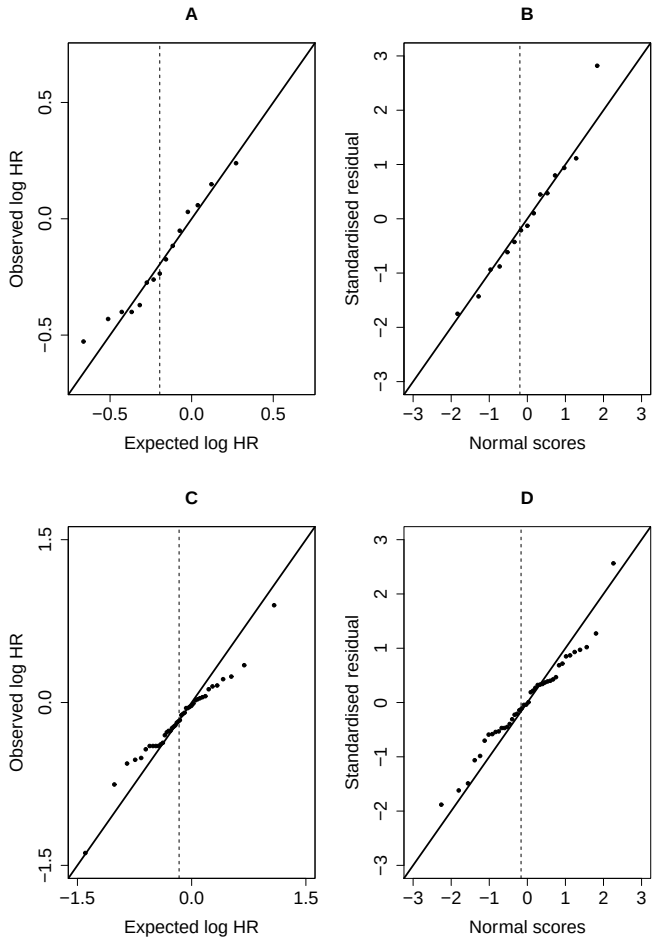

Figure 2: Observed and expected treatment differences from the largest 15 (Panel A and B) and 42 (Panel C and D) countries in the PLATO study. Panels A and C use absolute treatment effects whereas Panel B and C use the standardised weighted least squares residuals.

the P_E values restricted to the largest countries, as displayed in Panels A–C of Figure 3.

As a supplement to Figure 3, in Figure 4 we have displayed the observed and expected range of country-specific treatment effects for analyses restricted to the largest R countries, where R ranges from 10 through 42. It is clear from Figure 4 that regardless of whether the expected range of treatment effects is restricted to just the very large countries, or whether it includes the smaller countries with larger expected variation, the observed range of treatment effects is always consistent with the expected range.

4.5 Countries favouring the control

One feature that often causes concern in MRCTs with an overall experimental treatment benefit, is the occurrence of inconsistent country-specific treatment effects; that is, one or more country-specific treatment effects in the direction favouring the control

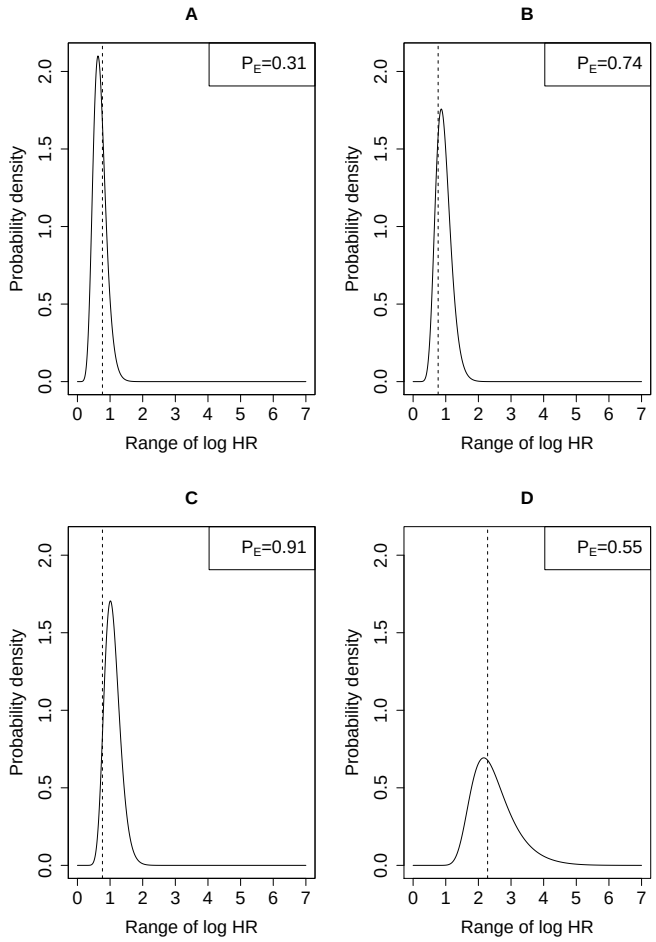

Figure 3: Probability density of the treatment effect range for the largest 10 (Panel A), 15 (Panel B), 20 (Panel C) and 42 (Panel D) countries in the PLATO study. The dotted line denotes the observed range.

treatment. This was certainly a concern in PLATO, particularly because one of these countries was the USA. In a study with as many countries as PLATO and a moderate overall treatment benefit, it is virtually certain that at least one country will have a treatment effect favouring the control, even if the treatment effect is homogeneous across countries. However, PLATO had 12 countries out 42 with treatment effects favouring the control, and when restricted to the largest countries, had 7 inconsistent effects out the largest 20 countries, 4 inconsistent effects out of the largest 15 countries, and 3 inconsistent effects out the largest 10 countries. These numbers of inconsistent countries may seem large, but when benchmarked against the probability distribution of the number of treatment effects favouring the control, as described in Section 3.3, it can be seen that they are not unusually large. Figure 5 displays these distributions, together with the observed numbers of inconsistent countries, and the the summary measure P_E which is the probability of an observation as least as extreme as the one observed. With $P_E = 0.72$ for the overall analysis in Panel D of Figure 5, it can be seen that an observation of 12 or more inconsistent countries is highly likely even under the assumption of treatment effect homogeneity. This conclusion is not altered by restrict-

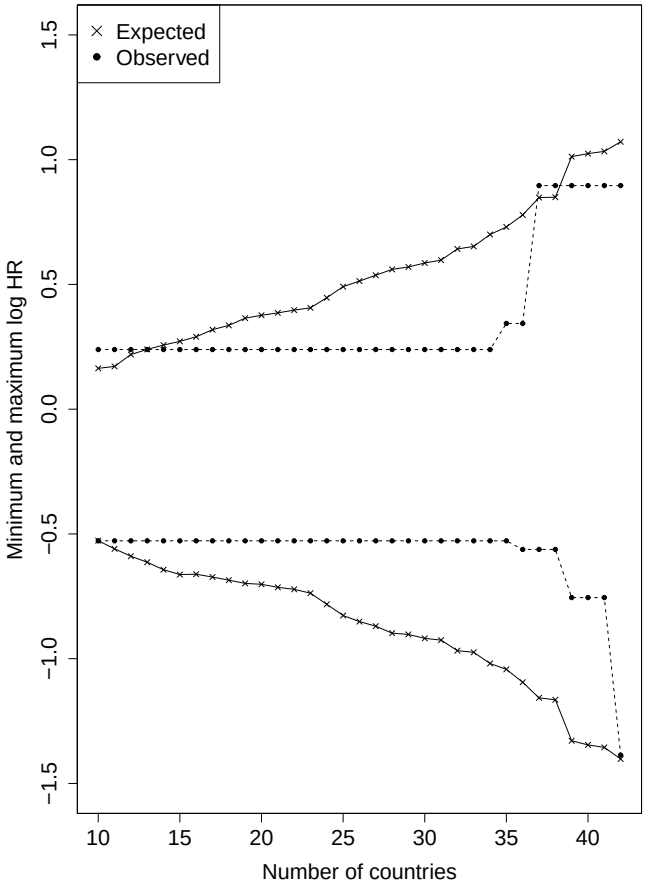

Figure 4: The range of observed and expected country-specific treatment effects in the PLATO study, restricting the analysis to the largest R countries, where R ranges from 10 to 42.

ing the analysis to the largest countries, as in Panels A–C of Figure 5, all of which also show that the observed number of inconsistent countries is not unusual relative to what would be expected by chance. Thus, these analyses suggest that any speculation about the causes of inconsistent country-specific treatment effects in PLATO, should acknowledge chance variation as a highly plausible explanation.

4.6 Conclusions

Despite all of the post hoc analysis and interpretation that the PLATO study has been subjected to, we conclude from our results that there is nothing particularly remarkable about the spread of treatment effects across countries. In a global study as large as the PLATO study, with over 40 countries, it is to be expected that wide variation in treatment effects will be observed. Consistent with earlier more limited analyses [22], our methods provide a suite of presentations suggesting that chance variation is a very plausible explanation for the spread of country-specific treatment effects observed in the PLATO study.

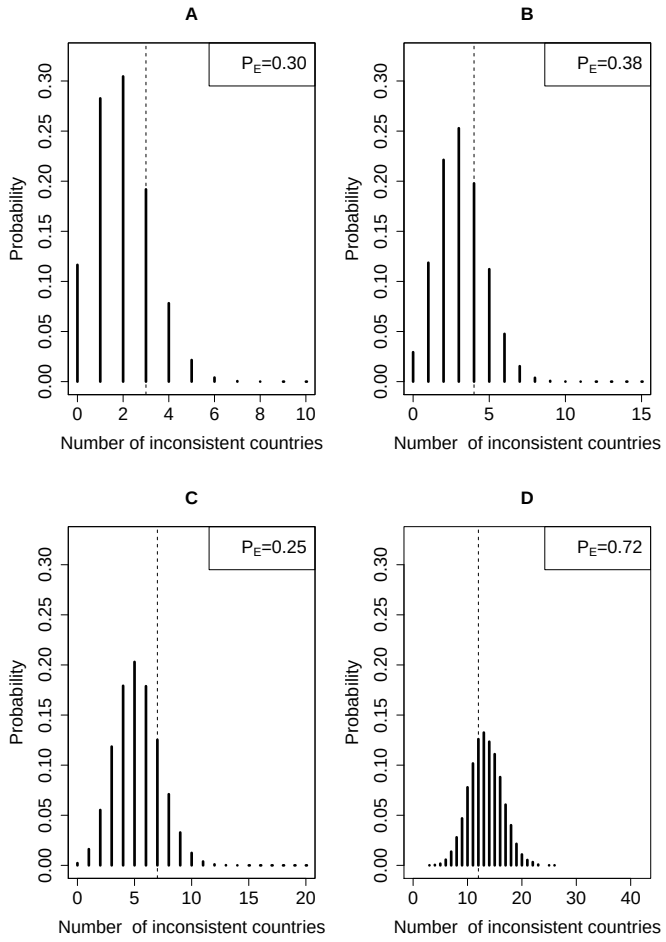

Figure 5: Probability distribution for the number of countries favouring the control for the largest 10 (Panel A), 15 (Panel B), 20 (Panel C) and 42 (Panel D) countries in the PLATO study. The dotted line denotes the observed value.

Finally, we note that our analyses were repeated to investigate how the various measures changed when a proportionally weighted overall standard error was used to estimate the s_r standard errors, as discussed in Section 3.5, instead of the individual country-specific standard errors used in the above analyses. It was found that there was very little difference between this approach and the approach presented in this section for the PLATO study.

5 Discussion

Assessment of heterogeneity of treatment effects between subgroups is a key element of clinical trial analysis. Recently, subgroup analysis of regional differences in MRCTs has become a prominent issue in the literature. In this paper we provide some new tools that aid interpretation of subgroup-specific treatment effects, and have illustrated these using a case study from a MRCT.

When a test of interaction is underpowered, and treatment effects are seemingly

different between subgroups, speculation may arise that there is heterogeneity of treatment effects that was not detected by the test of interaction. The approach we propose here is a non-inferential supplement to a formal test of interaction. A non-inferential approach has been suggested given that the same limitation of low power for a test of interaction will likely affect any new inferential technique one might develop to assess treatment effect heterogeneity. The suite of graphical tools introduced in this paper provide a multi-faceted visual assessment of the extent to which the observed treatment differences align with those that would be expected under an assumption of treatment effect homogeneity. That is, the intent is not to assess how these methods will perform under heterogeneity, but rather to quantify the potential extent of variation resulting from the play of chance under an assumption of homogeneity. Given the attention heterogeneity of treatment effects across regions has received in some MRCTs [14, 21, 23], our approach provides additional tools for evaluating the extent of chance variation expected in a MRCT, and can be used to benchmark expectations and pre-empt any over-interpretation. The graphical nature of our methods make it amenable for interpretation by all stakeholders including non-statisticians.

Treatment differences in typical clinical trial subgroups such as age and sex may present a plausible biological mechanism that explains the difference. However, treatment differences between regions are often more complex to understand because region is a composite of many variables that can potentially influence the outcomes of an intervention [4]. Thorough evaluation of potential treatment differences between regions at the design stage of a study is critical, and can assist with the interpretation of any apparent heterogeneity that emerges at the analysis stage.

Our methods differ from a recently published method by Chen et al. [13] in that we use the observed treatment differences whereas Chen et al. [13] use the standardised treatment differences as defined by the weighted least squares residuals. Although this difference may seem trivial, the results and their interpretation can be quite different as the ordering proposed by Chen et al. [13] depends on the relative magnitude of the departure of the region-specific treatment effect from the overall effect, compared with its standard error. We advocate the use of the observed treatment differences as these are required in practice for such activities as cost-effectiveness analyses and risk stratification in addition to the direct relevance they have for the physician and the patient.

In conclusion, our methods provide a non-inferential yet visually informative summary of the subgroup-specific variation in treatment effects that can arise as an artefact of chance. The appeal of these methods is their broad applicability, not just to global clinical trials as discussed here but also to other types of subgroup analysis, as well as the accessibility of the visual displays to all stakeholders including non-statisticians.

References

[1] Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications and interpretation. *Lancet* 2005; **365**:176-186.

[2] Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine – reporting of subgroup analyses in clinical trials. *The New England Journal of Medicine* 2007; **357;21**: 2189-2194.

[3] ICH. *Statistical Principles for Clinical Trials E9*, 1998. Available at: http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html

[4] Wittes J. Why is this subgroup different from all other subgroups? Thoughts on regional differences in randomized clinical trials. In: Fleming TR, Weir BS (eds), *Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials*, pp. 95–115. Springer, 2013. New York, USA.

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

[6] Marschner IC. Regional differences in multinational clinical trials: anticipating chance variation. *Clinical Trials* 2010; **7**:147–156.

[7] Chen J, Quan H, Binkowitz B, Ouyang SP, Tanaka Y, Li G, Menjoge S, Ibia E for the Consistency Workstream of the PhRMA MRCT Key Issue Team. Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. *Pharmaceutical Statistics* 2010; **9**:242-253.

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

[9] Chen J, Quan H, Gallo P, Menjoge S, Luo X, Tanaka Y, Li G, Ouyang SP, Binkowitz B, Ibia E, Talerico S, Ikeda K. Consistency of treatment effect across regions in multiregional clinical trials, part 1: design considerations. *Drug Information Journal* 2011; **45**:595-602.

[10] Buyse M, Squifflet P, Lucchesi KJ, Brune ML, Castaigne S, Rowe JM. Assessment of the consistency and robustness of results from a multicenter trial of remission maintenance therapy for acute myeloid leukemia. *Trials* 2011; **12**:86.

[11] Gallo P, Chen J, Quan H, Menjoge S, Luo X, Tanaka Y, Li G, Ouyang SP, Binkowitz B, Ibia E, Talerico S, Ikeda K. Consistency of treatment effect across regions in multiregional clinical trials, part 2: monitoring, reporting and interpretation. *Drug Information Journal* 2011; **45**:603-608.

[12] Ibia EO, Binkowitz B. Proceedings of the DIA Workshop on multiregional clinical trials, October 26-27, 2010. *Drug Information Journal* 2011; **45**:391-403.

[13] Chen J, Zheng H, Quan H, Li G, Gallo P, Ouyang SP, Binkowitz B, Ting N, Tanaka Y, Luo X, Ibia E and for the Society for Clinical Trials (SCT) Multi-Regional Clinical Trial Consistency Working Group. Graphical assessment of consistency in treatment effect among countries in multi-regional clinical trials. *Clinical Trials* 2013; DOI: 10.1177/1740774513500387 [epub ahead of print].

[14] Wallentin L, Becker RC, Budaj A, Cannon CP, Emanuelsson H, Held C, Horrow J, Husted S, James S, Katus H, Mahaffey KW, Scirica BM, Skene A, Steg PG, Storey RF, Harrington RA. Ticagrelor versus clopidogrel in patients with acute coronary syndromes. *New England Journal of Medicine* 2009; **361**:1045–1057.

[15] Schou IM, Marschner IC. Meta-analysis of clinical trials with early stopping: an investigation of potential bias. *Statistics in Medicine* 2013; DOI: 10.1002/sim.5893 [epub ahead of print].

[16] Li Z, Chuang-Stein C, Hoseyni C. The probability of observing negative subgroup results when the treatment effect is positive and homogeneous across all subgroups. *Drug Information Journal* 2007; **41**:47–56.

[17] Arnold BC, Balakrishnan N, Nagaraja HN. *A First Course in Order Statistics.* Society for Industrial and Applied Mathematics, 2008. Philadelphia, USA.

[18] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available at: www.R-project.org

[19] Balakrishnan N. Permanents, Order Statistics, Outliers, and Robustness. *Revista Matemática Complutense* 2007; **1**:7–107.

[20] FDA. *Ticagrelor Secondary Review*, 2010. Available at: http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/CardiovascularandRenalDrugsAdvisoryCommittee/UCM220192.pdf.

[21] Serebruany VL. Aspirin dose and ticagrelor benefit in PLATO: fact or fiction? *Cardiology* 2010; **117**:280–283.

[22] Buyse M, Marschner IC. Assessment of statistical heterogeneity in the PLATO trial. *Cardiology* 2011; **118**:138.

[23] Wedel H, DeMets D, Deedwania P, Fagerberg B, Goldstein S, Gottlieb S, Hjalmarson A, Kjekshus J, Waagstein F, Wikstrand J on behalf of the MERIT-HF Study Group. Challenges of subgroup analyses in multinational clinical trials: experiences from the MERIT-HF trial. *American Heart Journal* 2001; **142**:502–11.