

User manual for  
**Trial Sequential Analysis (TSA)**

**Kristian Thorlund, Janus Engstrøm, Jørn Wetterslev, Jesper Brok,  
Georgina Imberger, and Christian Gluud**

Copenhagen Trial Unit  
Centre for Clinical Intervention Research  
Department 3344, Rigshospitalet  
DK-2100 Copenhagen Ø  
Denmark  
Tel. +45 3545 7171 Fax +45 3545 7101  
E-mail: [tsa@ctu.dk](mailto:tsa@ctu.dk)

## Contents:

Disclaimer.....	1
Team member roles and contributions.....	2
Preface.....	3
1. Concepts and rationale behind trial sequential analysis.....	4
1.1. Random error in meta-analysis.....	4
1.2. Defining strength of evidence - information size.....	6
1.3. Testing for statistical significance before the information size has been reached.....	7
1.4. Testing for futility before the information size has been reached.....	9
1.5. Summary.....	10
2. Methodology behind TSA.....	12
2.1. Methods for pooling results from clinical trials.....	12
2.1.1. Effect measures for dichotomous and continuous data.....	12
2.1.2. General fixed-effect model and random-effects model setup.....	14
2.1.3. Approaches to random-effects model meta-analysis.....	16
2.1.4. Methods for handling zero-event trials.....	20
2.2. Adjusted significance testing and futility testing in cumulative meta-analysis.....	22
2.2.1. The information size required for a conclusive meta-analysis.....	24
2.2.2. The cumulative test statistic (Z-curve).....	34
2.2.3. Problems with significance testing in meta-analysis.....	35
2.2.4. The $\alpha$ -spending function and trial sequential monitoring boundaries.....	37
2.2.5. Adjusted confidence intervals following trial sequential analysis.....	44
2.2.6. The law of the iterated logarithm.....	47
2.2.7. The $\beta$ -spending function and futility boundaries.....	49
3. Installation and starting the TSA program.....	55
3.1. Prerequisites.....	55
3.2. Installation.....	55
3.3. Starting TSA.....	55
3.3.1. Why doesn't TSA start?.....	56
3.4. Starting RM5 Converter.....	57
3.4.1 Why doesn't RM5 start?.....	57
4. How to use TSA.....	58
4.1. Getting started.....	58
4.1.1. Creating a new meta-analysis.....	58
4.1.2. Saving a TSA file and opening an existing TSA file.....	60
4.1.3. Importing meta-analysis data from Review Manager v.5.....	60
4.2. Adding, editing, and deleting trials.....	64
4.2.1. Adding trials.....	64
4.2.2. Editing and deleting trials.....	66
4.3. Defining your meta-analysis settings.....	67
4.3.1. Choosing your association measure.....	67
4.3.2. Choosing your statistical model.....	68
4.3.3. Choosing a method for handling zero-event data.....	68
4.3.4. Choosing the type of confidence interval.....	69
4.4. Applying adjusted significance tests (applying TSA).....	70
4.4.1. Adding a significance test.....	71
4.4.2. Editing and deleting a significance test.....	78
4.4.3. Adding and loading significance test templates.....	79
4.4.4. Performing the significance test calculations.....	80
4.5. Graphical options for TSA.....	82
4.6. Exploring diversity across trials.....	86
5. TSA example applications.....	88
5.1. Datasets.....	88
5.2. Avoiding false positives.....	88
5.3. Confirming a positive result.....	90
5.3.2. Avoiding early overestimates.....	93
5.4. Testing for futility.....	95
5.5. Estimating the sample size of a new clinical trial.....	97
5.6. Other published trial sequential analysis applications.....	99

6.1. Effect measures for dichotomous and continuous data meta-analysis .....	102
6.2. Random-effects approaches .....	103
6.2.1. Formulas for the Biggerstaff-Tweedie method .....	103
6.3. Trial sequential analysis .....	103
6.3.1. Exaggerated type I error due to repeated significance testing .....	103
6.3.2. Alternative methods not implemented in the TSA software .....	104
7. List of abbreviations and statistical notation .....	108
7.1. General abbreviations .....	108
7.2. Statistical notation .....	108
7.2.1. Lower case letter symbols .....	108
7.2.2. Upper case letter symbols .....	109
7.2.3. Greek letter symbols .....	110

## **Disclaimer**

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

UNDER NO CIRCUMSTANCES AND UNDER NO LEGAL THEORY, WHETHER IN TORT, CONTRACT, OR OTHERWISE, SHALL COPENHAGEN TRIAL UNIT BE LIABLE TO YOU OR TO ANY OTHER PERSON FOR LOSS OF PROFITS, LOSS OF GOODWILL, OR ANY INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES, OR DAMAGES FOR GROSS NEGLIGENCE OF ANY CHARACTER INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOSS OF GOODWILL, WORK STOPPAGE, COMPUTER FAILURE OR MALFUNCTION, OR FOR ANY OTHER DAMAGE OR LOSS.

The Trial Sequential Analysis software (hereafter TSA) to which this manual refers is in Beta Release. Copenhagen Trial Unit has tested the TSA software extensively, but errors may still occur. Feedback is an important part of the process of correcting errors and implementing other changes, so we encourage you to tell us about your experiences with this software. To do so, please send your feedback to [tsa@ctu.dk](mailto:tsa@ctu.dk).

## **Team member roles and contributions**

TSA was developed at The Copenhagen Trial Unit, Copenhagen, Denmark. The team consisted of Kristian Thorlund (KT), Janus Engstrøm (JE), Jørn Wetterslev (JW), Jesper Brok (JB), Georgina Imberger (GI), and Christian Gluud (CG). The roles and contributions of each team member are outlined below:

Project manager: KT

Principal software application developer: JE.

Co-software application developers: KT, JW, JB, CG.

Statistical programmer: KT.

Internal beta-testers: JB, GI, JW, KT, CG.

Manual authors: KT (principal), GI, JW, JB, JE, CG.

Project supervisors: JW and CG.

## **Preface**

This manual provides a guide - both theoretical and practical - for the use of Copenhagen Trial Unit's Trial Sequential Analysis (TSA) software. Chapter 1 introduces the concepts and rationale, chapter 2 provides a technical overview of the implemented methodologies, and chapters 3-5 are practical chapters on how to install, use, and apply the software.

The TSA software can be downloaded at [www.ctu.dk/tsa](http://www.ctu.dk/tsa). You are welcome to use it in your analyses and publications of cumulative meta-analyses with proper reference to the software and some of our articles describing the methodology.

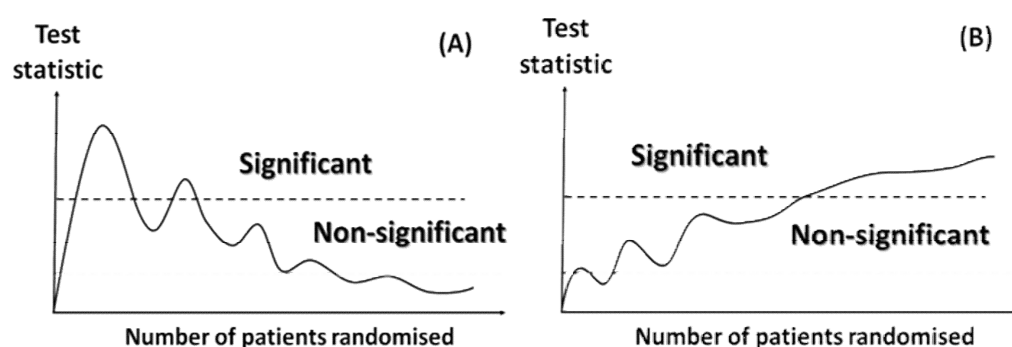
In case you need assistance with the TSA software, please contact us via email: [tsa@ctu.dk](mailto:tsa@ctu.dk).

## 1. Concepts and rationale behind trial sequential analysis

### 1.1. Random error in meta-analysis

Some 'positive' meta-analytic findings may be due to the play of chance (random error) rather than due to some underlying 'true' intervention effect.<sup>1-10</sup> Likewise, some neutral or 'negative' ('non-positive') meta-analytic findings may also represent a 'chance finding' due to lack of statistical power and precision.<sup>9-13</sup> These two types of errors are commonly known as false positive errors (or type I errors) and false negative errors (or type II errors). Meta-analyses are typically deemed 'positive' or 'negative' on the basis of some statistical test (test statistic), communicated with a P-value or with the corresponding confidence interval.

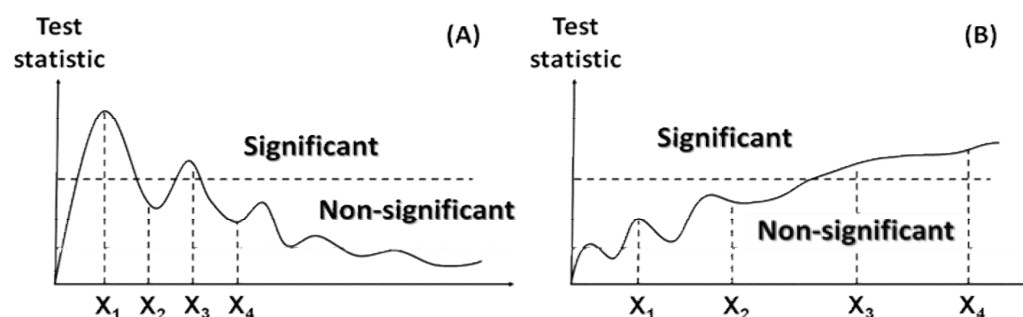
When a meta-analysis includes a small number of trials and a small number of patients, random errors can cause spurious findings.<sup>1;2;4-6;9;11;12;14;15</sup> Conversely, when there is a large number of patients, and when several trials have confirmed findings of previous trials, test statistics and intervention effect estimates will typically converge towards the 'truth'.<sup>1;2;4-6;9;11;12;14;15</sup> Figures 1(A) and 1(B) illustrate examples of such convergence in test statistics. In both situations, inferences about statistical significance are erroneous at certain early stages, but eventually converge to the 'true' side of statistical significance.



**Figure 1** Examples of convergence in test statistics as patients are included and followed to an outcome measure (e.g., death) in two randomised clinical trials A and B.

Random error and imprecision only cause problems if statistical tests (and intervention effect estimation) are employed at stages where the magnitude of the random error or imprecision is 'extreme enough' to yield spurious

statistical inferences. In figure 2(A), significance testing at times  $X_1$  and  $X_3$  would result in a false declaration of statistical significance (i.e., a false positive result), whereas significance testing at  $X_2$  and  $X_4$  would not. Thus, only at times  $X_1$  and  $X_3$  is the impact of random error 'extreme enough' to yield spurious statistically significant results. In figure 2(B), significance testing at  $X_1$  and  $X_2$  could have resulted in a false declaration that the interventions under investigation were not significantly different (i.e., a false negative result), whereas significance testing at  $X_3$  and  $X_4$  would not. Thus, only at times  $X_1$  and  $X_2$  is the imprecision of a magnitude that causes spurious absence of statistical significance.



**Figure 2** Examples of false positive and false negative statistical test results over time in two randomised clinical trials A and B.

The more statistical tests that are employed throughout the accumulation of additional data, the higher the likelihood of observing a false positive or false negative result. This phenomenon is commonly known as '*multiplicity due to repeated significance testing*',<sup>10;16-18</sup>

In meta-analysis it is important to minimize the risk of making a falsely positive or falsely negative conclusion.<sup>3</sup> Pooled intervention effects in meta-analysis are typically assessed on the basis of P-values. Meta-analysts must decide on the threshold at which a P-value is sufficiently small to justify a 'positive' conclusion. Below this threshold, a conclusion is considered statistically significant. At a given time, any threshold involves a trade-off between the risk of observing a false positive result (type I error) and the risk of observing a false negative result (type II error). For example, if the threshold for statistical significance in figure 2 (horizontal dashed line) had been moved up, the

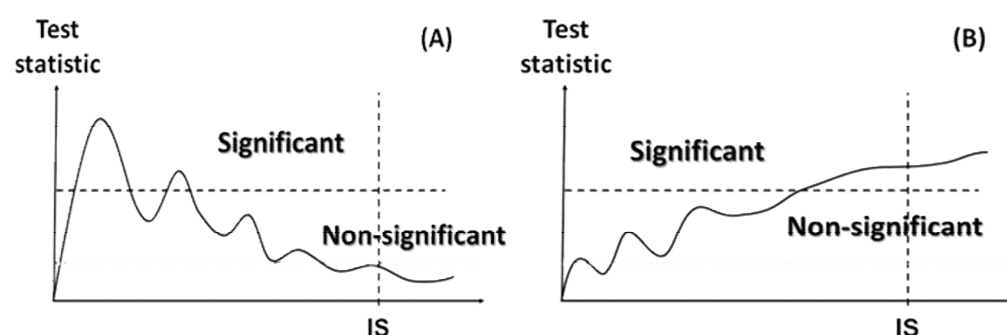


chance of observing a false positive result (figure 2(A)) would have diminished, while the risk of observing a false 'negative' result (figure 2(B)) would have increased. When conventional significance tests are performed at 'early' stages and/or at multiple times, these maximum risks are distorted (as illustrated in figure 2).<sup>16-18</sup> Thus, any inferences about statistical significance should be made in relation to the strength of the evidence. The strength of evidence should be measured using the accrued number of patients, observed number of events in the included trials, and the impact of multiplicity.<sup>1;2;4;6;10;19-21</sup>

## 1.2. Defining strength of evidence - information size

Meta-analyses of randomised trials increase the power and precision of the estimated intervention effects.<sup>13</sup> When all available trials are included, systematic reviews and meta-analyses are considered to be the best available evidence.<sup>13</sup> However, 'the best available evidence' may not be synonymous with 'sufficient evidence' or 'strong evidence'.<sup>1;2;4;6;11;12</sup>

In a single randomised trial with a binary outcome measure, we estimate the number of events and patients needed to allow for reliable statistical inference. That is, we perform a sample size calculation to ensure that a 'sufficient' number of events and patients are included.<sup>22</sup> A similar 'goal post' is needed for a meta-analysis.<sup>1;2;6;23</sup> This goal post has been referred to as the required meta-analysis information size (IS) or the optimum information size.<sup>1;2;4;6;11;12;14;15;19;23-25</sup> Figure 3 illustrates two typical meta-analytic scenarios A and B where the test statistic has stabilised after the required information size has been reached.



**Figure 3** Examples of how the required information size ensures reliable significance tests in two cumulative meta-analyses A and B.

A sample size calculation in a single trial is typically based on the expected control event proportion, the expected relative risk reduction of the experimental intervention, and the desired maximum risk of both type I error and type II error.<sup>26</sup> In a meta-analysis, there is likely heterogeneity across included trial populations, interventions, and methods. Meta-analysis sample size considerations need to be adjusted - that is, increased - in order to allow for the variance introduced by this heterogeneity.<sup>4;6;11;12;23</sup> Such adjustments are analogous to adjustments for variation across centres in a multi-centre trial.<sup>4;6;23</sup>

Conventional meta-analysis methods, such as those available in Review Manager v.5.1,<sup>27</sup> do not take into account the amount of the available evidence.<sup>13</sup> Instead, the reliability of a statistically significant intervention effect is commonly taken for granted, irrespective of the accrued number of events and patients. Conversely, intervention effects that are not statistically significant are commonly not considered reliable. Rather, it is assumed that 'more evidence is needed'.<sup>28</sup>

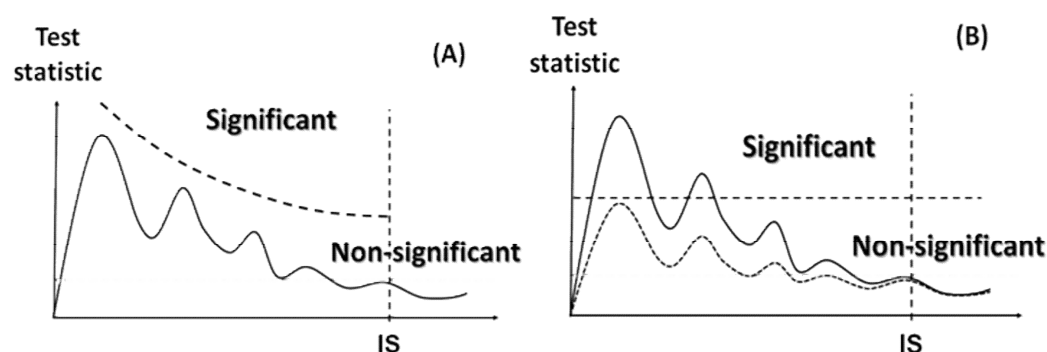
Empirical evidence suggests that intervention effects and P-values based on a limited number of events and patients are often not reliable.<sup>1;2;4-6;9;11;12;29</sup> About 25% of conventional meta-analyses that include a small number of events and patients may falsely declare the estimated intervention effects as statistically significant.<sup>4;5</sup> Empirical evidence also shows that large pooled intervention effects observed in early positive meta-analyses tend to dissipate as more evidence is accumulated.<sup>4;5;9</sup>

### **1.3. Testing for statistical significance before the information size has been reached**

The aim of a meta-analysis is to identify the benefit or harm of an intervention as early and as reliable as possible.<sup>4;11-13;20</sup> Therefore, meta-analyses are commonly updated when new trials are published. For example, Cochrane systematic review authors are required to update their systematic reviews at least every second year.<sup>13</sup> When meta-analyses are updated, they are repeatedly subjected to the significance testing over time. In randomised

clinical trials, repeated significance testing on accumulating data is known to inflate the overall risk of type I error.<sup>30</sup> Simulation studies suggest that if repeated significance testing is done in meta-analyses and P-values smaller than 0.05 are considered to be evidence of 'statistical significance', then the actual risk of type I error will be between 10% and 30%.<sup>7;8;10;31</sup> When decisions made accordingly to implement the intervention as a treatment, this means that between 1 and 3 out of 10 treatments decisions are likely inappropriate.

To deal with this problem, one can adjust the thresholds for which results are considered statistically significant and which results are not.<sup>1;2;4;6;11;12;14;15;24;25</sup> Alternatively, one can penalise the test statistic according to the strength of evidence and the number of performed significance tests (the 'law of the iterated logarithm').<sup>7;8</sup> The TSA software provides methods for both approaches, each building on theorems from advanced probability theory. The first approach uses methodology developed for repeated significance testing in randomised clinical trials (i.e., statistical monitoring boundaries).<sup>4;6;11;12</sup> The second approach penalizes - that is, decreases - the test statistic according to the strength of information available in the meta-analysis and the number of performed significance tests.<sup>7;8</sup>

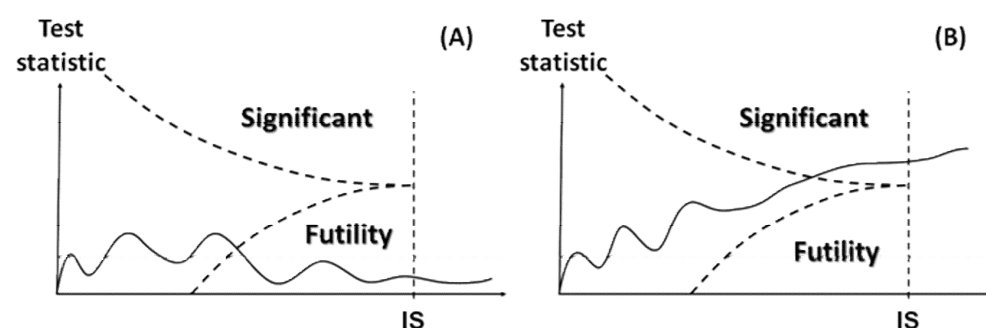


**Figure 4** Examples of significance threshold adjustment (stipulated monitoring boundaries) (A) and penalised test statistic (stipulated) (B) to avoid false positive statistical test results in two cumulative meta-analyses A and B.

Figure 4(A) illustrates an example of a meta-analysis scenario where a false positive result is avoided by adjusting the threshold for statistical significance by employing monitoring boundaries. Figure 4(B) illustrates an example where a false positive result is avoided by appropriately penalizing the test statistic.

#### 1.4. Testing for futility before the information size has been reached

It is also possible to use the TSA software to assess when an intervention is unlikely to have some anticipated effect. Or, in a clinical context, to assess when an intervention has an effect that is smaller than what would be considered minimally important to patients. Meta-analyses are often used to guide future research. Before embarking on future trials, investigators need to know an accurate summary of the current knowledge. If a meta-analysis has found that a given intervention has no (important) effect, investigators need to know whether this finding is due to lack of power or whether the intervention is likely to have no effect. Using conventional thinking, a finding of 'no effect' is considered to be due to lack of power until an appropriate information size has been reached. In some situations, however, we may be able to conclude earlier that a treatment effect is unlikely to be as large as anticipated, and thus, prevent trial investigators from spending resources on unnecessary further trials. Of course, the size of the anticipated intervention effect can be reconsidered and further research may be designed to investigate a smaller effect size.



**Figure 5** Examples of futility boundaries where the experimental intervention is not superior to the control intervention (and too many trials may have been conducted) (A) and where the experimental intervention is statistically significantly superior to the control intervention (and too many trials may have been conducted) (B).

TSA provides a technique for finding a conclusion of no effect as early as possible. 'Futility boundaries', which were originally developed for interim analysis in randomised clinical trials, are constructed and used to provide a threshold for 'no effect'.<sup>30</sup>

If the experimental intervention is truly superior to the control intervention, one would expect the test statistic to fluctuate around some upward sloping straight line, eventually yielding statistical significance (when the meta-analysis is sufficiently powered). If a meta-analysis of a truly effective experimental intervention includes only a small number of events and patients, the likelihood of obtaining a statistically significant result is low due to lack of power. However, as more evidence is accumulated, the risk of getting a chance negative finding decreases. Futility boundaries are a set of thresholds that reflect the uncertainty of obtaining a chance negative finding in relation to the strength of the available evidence (e.g., the accumulated number of patients). Above the thresholds, the test statistic may not have yielded statistical significance due to lack of power, but there is still a chance that a statistically significant effect will be found before the meta-analysis surpasses the IS. Below the threshold, the test statistic is so low that the likelihood of a significantly significant effect being found becomes negligible. In the latter case, further randomisation of patients is futile; the intervention does not possess the postulated effect.

Figure 5(A) illustrates an example where the experimental intervention is not superior to the control intervention. The test statistic crosses the futility boundaries (the upward sloping concave curve) before the required information size is surpassed. Figure 5(B) illustrates an example where the experimental intervention is statistically significantly superior to the control intervention. In this example, the test statistic stays above the futility curve (because there is an underlying effect) and eventually yields statistical significance.

### **1.5. Summary**

Trial sequential analysis (TSA) is a methodology that uses a combination of techniques. The evidence required is quantified, providing a value for the required IS. The thresholds for statistical significance are adjusted and these modifications are done according to the quantified strength of evidence and the impact of multiplicity.<sup>4;6;11;12</sup> Thresholds for futility can also be constructed, using a similar statistical framework.

In summary, TSA can provide an IS, a threshold for a statistically significant treatment effect, and the threshold for futility. Conclusions made using TSA show the potential to be more reliable than those using traditional meta-analysis techniques. Empirical evidence suggests that the information size considerations and adjusted significance thresholds may eliminate early false positive findings due to imprecision and repeated significance testing in meta-analyses.<sup>4;6;11;12</sup>

Alternatively, one can penalise the test statistic according to the strength of evidence and the number of performed significance tests (the 'law of the iterated logarithm').<sup>7;8</sup> Simulation studies have demonstrated that penalizing test statistics may allow for good control of the type I error in meta-analyses.<sup>7;8</sup>

The following manual provides a guide - both theoretical and practical - for the use of Copenhagen Trial Unit's TSA software. Chapter 2 provides a technical (intermediate level) overview of all the methodologies incorporated in the TSA software. Chapters 3-5 are practical chapters on how to install, use, and apply the TSA software.

## 2. Methodology behind TSA

TSA combines conventional meta-analysis methodology with meta-analytic sample size considerations (i.e., required information size) and methods already developed for repeated significance testing on accumulating data in randomised clinical trials.<sup>1;2;4;6;11;12</sup> In chapter 2, we first describe the meta-analysis methodology used to pool data from a number of trials. The description in section 2.1 covers effect measures for dichotomous and continuous data, statistical meta-analysis models (the fixed-effect model and some variants of the random-effects model), and methods for handling zero-event data. In section 2.2, we describe the methods for adjusting significance when there is an increased risk of random error (due to weak evidence and repeated significance testing). We do not describe the more advanced part of this methodology in detail. Rather, this chapter is intended to provide users with an intermediate level conceptual understanding of the issues addressed in chapter 1.

### 2.1. Methods for pooling results from clinical trials

#### 2.1.1. *Effect measures for dichotomous and continuous data*

The TSA program facilitates meta-analysis of dichotomous (binary) data and of continuous data. Dichotomous data are data that is defined by one of two categories (e.g., death or survival). Continuous data are data that is measured on a numerical scale (e.g., blood pressure or quality-of-life scores). For each type of data, there are various measures available for comparing the effectiveness of an intervention of interest.<sup>13</sup>

#### *Dichotomous data effect measures*

Assume we have  $k$  independent trials comparing two interventions (intervention A vs. intervention B) with a dichotomous outcome. Such trials will (typically) report the number of observed events (e.g., deaths) in the two intervention groups,  $e_A$  and  $e_B$ , and the total number of participants,  $n_A$  and  $n_B$ , in the two intervention groups. For dichotomous data, the intervention effect between the two interventions can be measured as risk difference (RD),

relative risk (RR), or odds ratio (OR).<sup>13</sup> Intervention effect estimates based on these measures are calculated using the following formulas:

$$RD = \frac{e_A}{n_A} - \frac{e_B}{n_B}$$

$$RR = \frac{(e_A / n_A)}{(e_B / n_B)}$$

$$OR = \frac{e_A / (n_B - e_B)}{e_B / (n_A - e_A)}$$

Relative risk ratios and odds ratios will typically be expressed on the log-scale because the log transformation induces certain desirable statistical properties (such as symmetry and approximate normality).<sup>13</sup> Standard errors, variances, and weights of 'ratio intervention effects' are therefore also obtained on the log-scale. The formulas for the standard errors of the RD, log(RR), and log(OR) are provided in appendix 6.1.

When the event proportions in the two groups are low (rare-event data), a preferred alternative to the odds ratio is the *Peto's odds ratio*.<sup>13</sup> This odds ratio is calculated with the formula:

$$OR_{Peto} = \exp((e_A - E(e_A)) / v)$$

Where  $E(e_A)$  is the expected number of events in intervention group A, and  $v$  is the (hypergeometric) variance of  $e_A$ . The formulas for  $E(e_A)$  and  $v$  are provided in appendix 6.1.

### *Continuous data effect measures*

Assume we have  $k$  independent trials comparing two interventions (intervention A vs. intervention B) with a continuous outcome. Such trials often report the mean response (e.g., mean quality of life score) in the two intervention groups,  $m_A$  and  $m_B$ , the standard deviations of the two



intervention group mean responses,  $sd_A$  and  $sd_B$ , and the total number of participants in the two intervention groups,  $n_A$  and  $n_B$ . When the mean response is measured on the same scale for all trials, comparative effectiveness is measured with the mean difference (MD), which is given by  $m_A - m_B$ . The standard error of the mean difference is given by

$$SE(MD) = \sqrt{\frac{sd_A^2}{n_A} + \frac{sd_B^2}{n_B}}$$

When the mean response is not measured on the same scale, mean responses can be *standardised* to the same scale, allowing for pooling across trials.<sup>11</sup> The conventional approach is to divide the mean response in each trial by its estimated standard deviation, thus providing an estimate of effect measured in *standard deviation units*. Mean differences divided by their standard deviation are referred to as standardised mean differences (SMD).<sup>13</sup>

**The TSA program does not facilitate meta-analysis of SMDs. Adjusted significance testing for SMD meta-analysis would require information size calculation be calculated on the basis of expected mean differences reported in standard deviation units. This effect measure does not resonate well with most clinicians and is therefore prone to produce unrealistic information size requirements.**

### **2.1.2. General fixed-effect model and random-effects model setup**

Assume we have  $k$  independent trials. Let  $Y_i$  be the observed intervention effect in the  $i$ -th trial. For dichotomous data meta-analysis,  $Y_i$  will either be the estimated risk difference, the log relative risk, the log odds ratio, or the log of Peto's odds ratio for the  $i$ -th trial. For continuous data meta-analysis,  $Y_i$  will be the estimated mean difference for the  $i$ -th trial. Let  $\mu_i$  be the true effect of the  $i$ -th trial and let  $\mu$  be the true underlying intervention effect (for the entire meta-analysis population). Let  $\sigma_i^2$  denote the variance (sampling error) of the observed intervention effect in the  $i$ -th trial.

In the fixed-effect model, the characteristics of the included trials (patient inclusion and exclusion criteria, administered variants of the intervention, study design, methodological quality, length of follow-up, etc.) are assumed to be similar.<sup>13</sup> This is formulated mathematically as  $\mu_1 = \mu_2 = \dots = \mu_k = \mu$ . The observed intervention effects of the individual trials are then assumed to satisfy the distributional relationship  $Y_i \sim N(\mu, \sigma_i^2)$ . The weight of a trial,  $w_i$ , is defined as the reciprocal of the trial variance, and hence, the trial weights, in a fixed-effect model, become  $w_i = \sigma_i^{-2}$ . The pooled intervention effect,  $\hat{\mu}$ , is obtained as a weighted average of the observed intervention effects of the individual trials

$$\hat{\mu} = \frac{\sum w_i Y_i}{\sum w_i}$$

and has variance

$$Var(\hat{\mu}) = \frac{1}{\sum w_i}$$

In the random-effects model, the intervention effects are assumed to vary across trials, but with an underlying true effect,  $\mu$ . Letting  $\tau^2$  denote the between-trial variance, the random-effects model is defined as follows

$$Y_i = \mu_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_i^2)$$

$$\mu_i = \mu + E_i, \quad E_i \sim N(0, \tau^2)$$

Where  $\varepsilon_i$  is the residual (sampling) error for trial  $i$ , and  $E_i$  is the difference between the 'true' overall effect and the 'true' underlying trial effect. Collapsing the hierarchical structure in the above equations,  $Y_i$  can be assumed to satisfy the distributional relationship  $Y_i \sim N(\mu, \sigma_i^2 + \tau^2)$ . Again, the trial weights are defined as the reciprocal of the variance, and so the trial weights in a random-effects model become  $w_i^* = (\sigma_i^2 + \tau^2)^{-1}$ . The meta-analysed intervention effect,  $\hat{\mu}$ , is obtained as a weighted average of the observed intervention effects of the individual trials.

$$\hat{\mu} = \frac{\sum w_i^* Y_i}{\sum w_i^*}$$

and has variance

$$Var(\hat{\mu}) = \frac{1}{\sum w_i^*}$$

Statistical significance testing is performed with the Wald-type test statistic, which is equal to the meta-analysed intervention effect (log scale for relative risks and odds ratios) divided by its standard error:

$$Z = \frac{\hat{\mu}}{\sqrt{Var(\hat{\mu})}}$$

This test statistic is typically referred to as the Z-statistic or the Z-value. Under the assumption that the two investigated interventions do not differ the Z-value will approximately follow a standard normal distribution (a normal distribution with mean 0 and standard deviation 1). This assumption is also referred to as the *null hypothesis* and is denoted  $H_0$ . The corresponding two-sided P-value can be obtained using the following formula:

$$P = 2 \cdot (1 - \Phi(|Z|))$$

where  $|Z|$  denotes the absolute value of the Z-value and  $\Phi$  denotes the cumulative standard normal probability distribution function.<sup>13</sup> The P-value is the probability of observing a Z-value at least as 'extreme' as the one observed due to the play of chance. The smaller the P-value, the smaller is the likelihood that the difference observed between two intervention groups is simply a chance finding, and thus, the larger is the likelihood that the observed difference was caused by some underlying 'true' treatment effect.

### **2.1.3. Approaches to random-effects model meta-analysis**

As explained above, the random-effects model attempts to include a quantification of the variation across trials.<sup>13</sup> The common approach is to

estimate the between-trial variance,  $\tau^2$ , with some between-trial variance estimator.<sup>13</sup>

#### *The DerSimonian-Laird method*

The between-trial variance estimator which has been used most commonly in meta-analytic practice (and is the only option in The Cochrane Collaboration's *Review Manager* software) is the estimator proposed by DerSimonian and Laird (DL).<sup>13;27;32</sup> The DL estimator takes the form

$$\tau_{DL}^2 = \max(0, (Q - k + 1) / (S_1 - (S_2 / S_1)))$$

Where  $Q$  is the Cochran homogeneity test statistic given by  $Q = \sum w_i (Y_i - \hat{\mu})^2$ , where  $S_r = \sum w_i^r$ , for  $r = 1, 2$ , and where  $k$  is the number of trials included in the meta-analysis.<sup>13;32</sup>

Because the DL estimator is prone to underestimate the between-trial variance,<sup>33-40</sup> we have included two alternative random-effects model approaches – the Sidik and Jonkman (SJ) and the Biggerstaff and Tweedie (BT) methods - in the TSA software.<sup>33;34;41</sup>

#### *The Sidik-Jonkman (SJ) method*

The SJ random-effects model uses a simple (non-iterative) estimator of the between-trial variance based on a re-parametrisation of the total variance of the observed intervention effect estimates  $Y_i$ .<sup>33;34</sup> It is given by the expression:

$$\tau_{SJ}^2 = \sum v_i (Y_i - \mu_0)^2 / (k-1)$$

where  $v_i = r_i + 1$ ,  $r_i = \sigma_i^2 / \tau_0^2$ , and  $\tau_0^2$  is an initial estimate of the between-trial variance, which can be defined, for example, as

$$\tau_0^2 = \sum (Y_i - \mu_{uw})^2 / k$$

$\mu_{uw}$  being the unweighted mean of the observed trial effect estimates, and  $\mu_0$  being the weighted random-effects estimate using  $\tau_0^2$  as the estimate for the

between-trial variance. Simulation studies have demonstrated that the SJ estimator provides less downward-biased estimates of the between-trial variance than the DL estimator.<sup>34;37</sup> That is, the SJ method is less likely to under-estimate the heterogeneity between trials. This is particularly the case for meta-analysis data that incur moderate or substantial heterogeneity. Confidence intervals based on the SJ estimator have coverage close to the desired level (e.g., 95% confidence intervals will contain the true effect in approximately 95% of all meta-analyses).<sup>34;37</sup> In contrast, the commonly reported coverage of confidence intervals based on the DL estimator is often below the desired level.<sup>33;35-38</sup> For example, many simulation studies that have investigated the coverage of DL-based 95% confidence intervals have found an actual coverage of 80%-92%.<sup>34;37</sup> The size of these confidence intervals is equivalent to a false positive proportion of 8% to 20%, which is clearly larger than the conventionally accepted 5%.

#### *The Biggerstaff-Tweedie method*

Because most meta-analyses contain only a limited number of trials, between-trial variance estimation is often subject to random error.<sup>41</sup> Incorporating the uncertainty of estimating the between-trial variance in the random-effects model may therefore be warranted. Biggerstaff and Tweedie (BT) proposed a method to achieve such incorporation.<sup>41</sup> They derived an approximate probability distribution,  $f_{DL}$ , for the DL estimate of  $\tau^2$ . Defining the trial weights as  $w_i(t) = (\sigma_i^2 + t)^{-1}$ , where  $t$  is a variable that can assume all possible values for  $\tau^2$ , they utilised  $f_{DL}$  and obtained trial weights that take the uncertainty of estimating  $\tau^2$  into account. This generally creates a weighting scheme which, relative to the DL approach, attributes more weight to larger trials and less weight to smaller trials. Biggerstaff and Tweedie also proposed an adjusted formula for the variance of the meta-analysed intervention effect, thus facilitating adjusted confidence intervals (see appendix, section 6.2.1).

#### *Which random-effects approach may be best?*

The SJ and BT approaches both offer relative merits over the DL approach. However, these methods have their own limitations and are unlikely to be superior in all cases. The SJ estimator may overestimate the between-trial

variance in meta-analyses with mild heterogeneity, thus producing artificially wide confidence intervals.<sup>34;37</sup> The BT approach has been shown to provide similar coverage as the confidence intervals from the DL approach in meta-analyses with small, unbiased trials.<sup>35</sup> However, when the included trials differ in size and some small trials are biased, the BT approach will put appropriately high weights on the larger trials while still accounting for heterogeneity. This point is important because a common critique of the DL random-effects model is that small trials are often assigned artificially large weights in heterogeneous meta-analyses. A commonly applied, and unsatisfactory, solution is to use the fixed-effect model instead. By doing so, the pooled estimate may incur less bias from the inappropriate weighting scheme, but the confidence intervals will also be artificially narrow because they do not account for heterogeneity. The BT approach mitigates the bias incurred from inappropriate random-effects model weighting while still accounting for heterogeneity.

The choice of random-effects model should involve a sensitivity analysis comparing each approach. If the DL, SJ, and BT approaches all yield similar statistical inferences (i.e., point estimates and confidence intervals), it would be reasonable to use the DL approach and have confidence that the estimation of between trial variance is reliable.

If two (or all) of the three approaches differ, one should carry out meta-analysis with both (or all) approaches and consider the results according to the underlying properties of each approach. For example, if the DL and SJ approaches produce different results, two possible explanations should be considered: 1) the meta-analysis is subject to moderate or substantial heterogeneity and the DL estimator therefore underestimates the between-trial variance and yields artificially narrow confidence interval; and 2) the meta-analysis is subject to mild heterogeneity and the SJ estimator therefore overestimates the between-trial variance and yields artificially wide confidence intervals. In this situation, one should then carry out meta-analyses with the two approaches and consider the implications of each of the two scenarios being 'true'.

#### **2.1.4. Methods for handling zero-event trials**

In dichotomous trials, the outcome of interest may be rare. For example, the occurrence of heart disease from the use of hormone replacement therapy is very low.<sup>42</sup> Sometimes there are zero outcome events recorded in a group. In this situation, ratio effect measures (RR and OR) will not give meaningful estimates of the intervention effect.<sup>42</sup> One solution for this problem is to add some constant(s) to the number of events and non-events in both intervention groups.<sup>42</sup> This approach is known as *continuity correction*.<sup>42</sup> Several approaches to continuity correction have been proposed in the meta-analytic literature.

##### *Constant continuity correction*

The constant continuity correction is a simple method and is the most commonly used in the meta-analytic literature.<sup>42</sup> The method involves adding a *continuity correction factor* (a constant) to the number of events and non-events in each intervention group.

Group	Events	No Events	Total
Intervention	0	20	20
Control	5	20	25

**Table 1** Example of a zero-event trial

Consider the zero-event trial example in table 1. If, for example, the constant continuity correction method uses a correction factor of 0.5, the number of events in the intervention group becomes  $0+0.5=0.5$ , the number of non-events in the intervention group becomes  $20+0.5=20.5$ , the number of events in the control group becomes  $5+0.5=5.5$ , and the number of non-events in the control group becomes  $20+0.5=20.5$ . Because the total number of patients is the number of events plus the number of non-events, the total number of patients (after constant continuity correction with the constant 0.5) becomes  $20.5+0.5=21$  in the intervention group and  $20.5+5.5=26$  in the control group.

If, for example, a correction factor of 0.1 is used, the number of events and total number of patients (after continuity correction) would then be 0.1 and 20.2 in the intervention group and 5.1 and 25.2 in control group.

Review Manager Version 5 uses constant continuity correction with the constant 0.5.<sup>13;27</sup> Simulation studies have demonstrated problems with the use of this constant; it yields inaccurate estimates when the randomisation ratio is not 1:1, and it produces confidence intervals that are too narrow.<sup>42</sup>

#### *Reciprocal of opposite intervention group continuity correction*

Another potential continuity correction method is to add the reciprocal of the total number of patients in the opposite intervention group to the number of events and non-events.<sup>42</sup> This type of continuity correction is also commonly referred to as 'treatment arm' continuity correction.<sup>42</sup> In the example in table 1, the correction factor for the intervention group would be  $1/25=0.04$ , and the correction factor for the control group would be  $1/20=0.05$ . This continuity correction method yields 0.04 events and 20.04 patients in the intervention group and 5.05 events and 25.05 patients in the control group.

#### *Empirical continuity correction*

Both the constant continuity correction method and the 'treatment arm' continuity correction method pull the intervention effect estimates towards 'the null effect' (i.e., towards 0 for risk differences and toward 1 for ratio measures).<sup>42</sup> An alternative continuity correction is the *empirical* continuity correction which pulls the intervention effect estimate towards the meta-analysed effect.<sup>42</sup> For example, let  $\hat{\theta}$  be the odds ratio of the meta-analysis that does not include the zero-event trials, and let  $R$  be the randomisation ratio in the trial that needs continuity correction. The continuity correction factor for the intervention group,  $CF_I$ , and the continuity correction for the control group,  $CF_C$ , can be approximated with the following formulas:



$$CF_I = \frac{R}{R + \hat{\theta}} \cdot C$$

$$CF_C = \frac{\hat{\theta}}{R + \hat{\theta}} \cdot C$$

under the restriction that the two continuity corrections add up to some constant  $C$ .<sup>42</sup>

## 2.2. Adjusted significance testing and futility testing in cumulative meta-analysis

Adjusted significance testing in cumulative meta-analysis has two goals: it must measure and account for the strength of the available evidence and it must control the risk of statistical errors (type I error and type II error) when repeated significance testing on accumulating data occurs.

Quantifying the strength of the available evidence necessitates the definition of a 'goal post'.<sup>1;2;4;6;11;12;23</sup> In the TSA programme (TSA), the strength of available evidence is measured, and considered, by calculating a required information size. This information size is analogous to the required sample size in a single randomised clinical trial.<sup>1;2;4;6;11;12;23</sup>

Controlling the risk of type I error involves an alteration in the way we measure statistical significance. If a meta-analysis is subjected to significance testing before it has surpassed its required information size, the threshold for statistical significance can be adjusted to account for the elevated risk of random error.<sup>1;2;4;6;11;12;23</sup> Alternatively, the test statistic itself can be penalised in congruence with the strength of the available evidence. TSA provides the option to use both of these approaches to control the type 1 error.

Controlling the risk of type II error before a meta-analysis surpasses its required information size involves setting up thresholds (rules) for when the experimental intervention can be deemed non-superior (and/or non-inferior) to the control intervention.

The methods for adjusting significance thresholds (i.e., controlling the type I error) build on methods introduced by Armitage and Pocock; these methods are referred to as 'group sequential analysis'.<sup>18;43;44</sup> In Armitage's and Pocock's group sequential analysis, it is necessary to know the approximate number of patients randomised between each *interim look* at the data.<sup>30</sup> In randomised clinical trials, interim looks on accumulating data are typically pre-planned and it is therefore possible to define known group sizes between each interim look.<sup>30</sup> In meta-analysis, an interim look at the data occurs when there is an update, adding data from new clinical trials. Updates in meta-analysis occur at an arbitrary pace, are seldom regular, and the number of added patients is varied and unpredictable. The methods proposed by Armitage and Pocock are therefore inapplicable for meta-analysis.

Lan and DeMets extended the methodology proposed by Armitage and Pocock, allowing for flexible, unplanned *interim analyses*. Lan and DeMets intended this methodology for repeated significance testing in a single randomised trial.<sup>16;17;30</sup> Because of the flexibility of the timing of interim looks, this methodology is applicable to meta-analysis. The Lan and DeMets approach is therefore the methodology used in TSA; it involves construction of monitoring boundaries that facilitate the definition of sensible thresholds for 'statistical significance' in meta-analysis.

Similarly, futility boundaries can be constructed, facilitating the definition of sensible thresholds for 'futility' in meta-analysis.<sup>30</sup> Sections 2.2.1. to 2.2.5. provide a description of the underlying methodology and theoretical considerations for these methods.

The methods for controlling for type II error are an extension of the Lan-DeMets methodology that allows for non-superiority and non-inferiority testing. That is, instead of constructing adjusted thresholds for statistical significance, the method constructs adjusted thresholds for non-superiority and non-inferiority (or *no difference*). Together, adjusted non-superiority and non-inferiority boundaries make up what is referred to as futility boundaries or

*inner wedge* boundaries. Sections 2.2.7. provides a description of the underlying methodology and theoretical considerations for this method.

As previously described, an alternative approach to the alteration of thresholds is to penalise the test statistic itself. The method for penalising the employed statistical tests is a relatively new approach, which builds on theorems from advanced probability theory. In particular, the technique uses the theorem known as '*the law of the iterated logarithm*'.<sup>7;8</sup> Sections 2.2.2 and 2.2.6 provide a description of the underlying methodology and theoretical considerations for this method.

### **2.2.1. The information size required for a conclusive meta-analysis**

Determining the required information size (e.g., the required number of patients) for a conclusive and reliable meta-analysis is a prerequisite for constructing adjusted thresholds for 'statistical significance' using TSA.<sup>1;2;4;6;11;12</sup> The levels of the thresholds must be constructed in accordance with the strength of evidence.<sup>1;2;4;6;11;12</sup> The statistical methodology underlying TSA is based on the assumption that data will accumulate until the required information size is surpassed.<sup>30</sup> For further explanation on this assumption, please refer to earlier methodological papers on this issue.<sup>16;17;30;43;44</sup>

#### *Conventional information size considerations*

It has been argued that the sample size required for a conclusive and reliable meta-analysis should be at least as large as the sample size required to detect a realistic intervention effect in a large, reasonably powered trial.<sup>1;2;4;6;11;12</sup> In line with this construct, the minimum required information size (number of patients) in a meta-analysis can be derived using the well-known formula:

$$IS_{Patients} = 2 \cdot (Z_{1-\alpha/2} + Z_{1-\beta})^2 \cdot 2 \cdot \sigma^2 / \delta^2 \quad (1)$$

where  $\alpha$  is the desired maximum risk of obtaining a false positive result (type I error) and  $\beta$  is the desired maximum risk of obtaining a false negative result (type II error), and where  $Z_{1-\alpha/2}$  and  $Z_{1-\beta}$  are the  $(1- \alpha/2)$  and  $(1- \beta)$  standard

normal distribution quantiles.<sup>1;2;4;6;11;12</sup> Note that the use of  $\alpha / 2$  instead of  $\alpha$  means that the information size is constructed assuming two-sided statistical testing. For binary data,  $\delta = P_C - P_E$  denotes an *a priori* estimate for a realistic or minimally important intervention effect ( $P_C$  and  $P_E$  being the proportion with an outcome in the control group and the in the intervention group, respectively), where  $\sigma^2 = P^* (1 - P^*)$ , which is the associated variance, and assuming  $P^* = (P_C + P_E) / 2$  (i.e., that the intervention and control groups are equal in size). For continuous data,  $\delta$  denotes an *a priori* estimate of the difference between means in the two intervention groups, and  $\sigma^2$  denotes the associated variance.

#### *Alternatives to accumulating number of patients*

In meta-analysis of binary data, the information and precision in a meta-analysis predominantly depends on the number of events or outcomes. One can therefore argue that in the context of meta-analysis information size considerations, the required number of events is a more appropriate measure than the required number of patients. Under the assumption that an equal number of patients are randomised to the two investigated interventions in all trials, the required number of events may be determined as follows:

$$IS_{Events} = P_C * IS/2 + P_E * IS/2$$

where  $IS_{Events}$  is the required number of events for a conclusive and reliable meta- analysis, and  $P_C$  and  $P_E$  are as defined in the previous paragraph.

The *statistical information* (Fischer information) is a statistical measure of the information contained in a data set (given some assumed statistical model).<sup>45</sup> In standard meta-analysis comparing two interventions, the statistical information is simply the reciprocal of the pooled variance.<sup>46</sup> In a meta-analysis, the statistical information is a theoretically advantageous measure because it combines three factors in one single measure: number of patients, number of events, and number of trials. This measure provides a simple approach to information size considerations in a meta-analysis. The meta-

analytical data are considered as analogous to accumulating data in a single trial and the required statistical information is given by:

$$IS_{Statistical} = (Z_{1-\alpha/2} + Z_{1-\beta})^2 / \delta^2$$

Where  $IS_{Statistical}$  is the actual attained statistical information in the meta-analysis,  $\alpha$  is the desired maximum risk of type I error,  $Z_{1-\alpha/2}$  is the standard normal  $(1 - \alpha/2)$  percentile,  $\beta$  is the desired maximum risk of type II error,  $Z_{1-\beta}$  is the standard normal  $(1 - \beta)$  percentile, and  $\delta$  is some pre-specified (minimally relevant) intervention effect.<sup>30;45</sup>

#### *The heterogeneity-adjustment factor*

Trials included in a meta-analysis often include patients from a wide span of population groups, use different regimens of an intervention, use different study designs, and vary in methodological quality (i.e., risk of bias or 'systematic error'). For all of these reasons, it is natural to expect an additional degree of variation in meta-analysis data compared to data from a single trial.<sup>13;47</sup> Such additional variation is referred to as heterogeneity (or between-trial variation).<sup>13;47</sup> Because increased variation can decrease the precision of results, information size considerations must incorporate all sources of variation in a meta-analysis, including heterogeneity.<sup>4;6;11;12</sup> One approach for incorporating heterogeneity in information size considerations is to multiply the required information size in a meta-analysis by some *heterogeneity-adjustment factor*.<sup>6;23</sup> Recently, a similar *heterogeneity-adjustment factor* has been proposed for estimating the sample size in a single clinical trial.<sup>48</sup>

The heterogeneity adjustment factor is conceptualised through the underlying assumptions that we make for our meta-analysis model. In the fixed-effect model, it is assumed that all included trials can be viewed as replicates of the same trial (with respect to design and conduct). Thus, the required information size for a fixed-effect meta-analysis to be conclusive may effectively be calculated in the same way as the required sample size for a single clinical trial. In the random-effects model, we assume that the included trials come from a distribution of possible trials (with respect to design and conduct). By

definition, the variance in a random-effects model is always greater than that in a fixed-effect model. A heterogeneity-adjustment factor must therefore account for the increase in variation that a meta-analysis incurs from going from the fixed-effect assumption to the random-effects assumption. An accurate adjustment can be achieved by making the heterogeneity-adjustment factor equal to the ratio of the total variance in a random-effects model meta-analysis and the total variance in a fixed-effect model meta-analysis.<sup>6;23</sup> The heterogeneity-adjustment factor is therefore always equal to or greater than 1. Letting  $IS_{Fixed}$  denote the required information size for a fixed-effect meta-analysis given by equation (1),  $v_R$  denote the total variance in the random-effects model meta-analysis, and  $v_F$  denote the total variance in the fixed-effect model meta-analysis, the heterogeneity-adjusted information size can be derived using the following formula:

$$IS_{Random} = \frac{v_R}{v_F} IS_{Fixed}$$

Given that the anticipated intervention effects in the fixed- ( $\delta_F$ ) and random-effects ( $\delta_R$ ) models are approximately equal (that is, given  $\delta_R = \delta_F$ ), it can be shown mathematically that in the special case where all trials in a meta-analysis are given the same weights, the heterogeneity-adjustment factor (AF) takes the form

$$AF = \frac{v_R}{v_F} = \frac{1}{1 - I^2}$$

Where  $I^2$  is the inconsistency factor commonly used to measure heterogeneity in a meta-analysis.<sup>47</sup>

It is important to remember that in any case where the trial weights are not equal, using  $I^2$  will lead to an underestimation of the adjustment factor, and thus, an underestimation of the required information size.<sup>23</sup> In this situation, we can define a measure of diversity ( $D^2$ ) as the quantity compelled to satisfy the equation:

$$AF = \frac{v_R}{v_F} = \frac{\sum_{i=1}^k w_i}{\sum_{i=1}^k w_i^*} = \frac{1}{1 - D^2}$$

where  $w_i$  denotes the trial weights in the fixed-effect model and  $w_i^*$  denotes the trial weights in the random-effects model. Solving the equation with respect to  $D^2$ , we get:

$$D^2 = \frac{v_R - v_F}{v_R} = 1 - \frac{v_F}{v_R} = 1 - \frac{\sum_{i=1}^k w_i^*}{\sum_{i=1}^k w_i} = \frac{\sum_{i=1}^k (w_i^{-1} + \tau^2)^{-1}}{\sum_{i=1}^k w_i}$$

where  $\tau^2$  denotes the between-trial variance. One advantageous property of the diversity measure,  $D^2$ , is that the above derivations are generalisable to any given meta-analysis model. Thus, if we wish to meta-analyse some trials using an alternative random-effects model with total variance  $v_R$ , the diversity measure and the corresponding adjustment factor simply take the expression:

$$D^2 = \frac{v_R - v_F}{v_R} \quad \text{and} \quad AF = \frac{v_R}{v_F}$$

Estimates of variability, and in particular between-trial variability, may be subject to both random error and bias.<sup>41;47;49;50</sup> For this reason, in some situations, using  $D^2$  or  $I^2$  based on the available data may be inappropriate. In meta-analyses that only include a limited number of trials (e.g., less than 10 trials), estimates of heterogeneity and the between-trial variance may be just as unreliable as intervention effect estimates from small randomised clinical trials (e.g., trials including less than 100 patients). When a meta-analysis is subject to time-lag bias (i.e., when trials, mostly with positive findings, have been published), the between-trial variance will typically be underestimated. This underestimation occurs because the 'early' set of included trials are likely to have yielded similar ('positive') intervention effect estimates.<sup>50</sup> Later meta-

analyses (updates) are likely to include more trials with neutral or even negative findings, in which cases the estimates of heterogeneity will be larger.

For meta-analyses with an expected small number of trials, we suggest that an a priori estimate about the anticipated degree of heterogeneity is made. If we let  $H$  denote a conceptual estimate of  $D^2$ , we can use the following formula in an *a priori* calculation:

$$AF = \frac{1}{1-H}$$

For example, if it is expected that a given meta-analysis will contain a mild degree of heterogeneity – based on what we know about the clinical topic, observed differences between the included trials, anticipated differences between current and future, and the scope of the review – one may choose to define  $H$  as 25%. In this case, the  $AF$  would be estimated at 1.33. If a moderate degree of heterogeneity is expected, one may choose to define  $H$  as 50%, and  $AF$  would then be estimated at 2.00. If major heterogeneity is expected, then  $H$  may become 75% and  $AF$  would be estimated to 4.00.

Because the expected degree of heterogeneity can be difficult to estimate when a meta-analysis only includes a few trials, we recommend that users of TSA conduct sensitivity analyses for this variable. For example, one could conceive minimum and maximum realistic or acceptable degrees of heterogeneity for a given meta-analysis. As an example, one could speculate that the minimum plausible degree of statistical heterogeneity would be 20%. One could also decide that if the statistical heterogeneity exceeds 60%, then subgroup effect measures, rather than estimating an overall pooled estimated treatment effect, would be more appropriate. In this case, the over-all meta-analysis would not be performed. In this example, one could use the average of the two,  $(60\%+20\%)/2=40\%$ , for the primary information size calculation, but acknowledge that the required information size may be as large as the one based on 60% heterogeneity adjustment or as low as the one based on 20% heterogeneity adjustment. As another example, one could conceive and



construct a number of 'best' and 'worst' case scenarios (whatever those might be) by adding 'imaginary' future trials to the current meta-analysis. This approach would allow one to assess the robustness and reliability of the  $D^2$  estimate and construct a spectrum of realistic or acceptable degrees of heterogeneity which could readily be utilized for sensitivity analysis.

*Estimating the control group event proportion and an anticipated intervention effect*

The estimation of the control group event proportion and an anticipated intervention effect are important determinants of the calculated required information size when doing TSA. Every effort should therefore be made to make these estimates as accurate and realistic as possible.

For binary data, control group event proportion can be estimated by using clinical experience and evidence from related areas. An *a priori* estimate of a realistic intervention effect is usually expressed as a relative risk reduction (RRR). When there is limited evidence available about the intervention under investigation, one can estimate a clinically relevant intervention effect by using clinical experience and evidence from related areas. An example can be found in a paper by Pogue and Yusuf, in which the control group event proportion,  $P_C$ , and an *a priori* RRR were based on experiences from related areas in cardiology.<sup>1;2</sup> Pogue and Yusuf applied information size considerations to two well-known meta-analyses in cardiology: 'Intravenous Streptokinase in Acute Myocardial Infarction' and 'Intravenous Magnesium in Acute Myocardial Infarction'. They hypothesized that for most major vascular outcomes, such as death, it may be realistic to expect 10% mortality in the control group. Pogue and Yusuf further considered an example of a theoretical intervention for preventing mortality post myocardial infarction. They noted that truly effective treatments for reducing the risk of major cardiovascular events, such as death, had previously yielded RRRs of 10%, 15%, or - at best - 20%.

For any given clinical question, a decision needs to be made about what values are appropriate for the  $P_C$  and RRR. The anticipated proportion of

events in the (experimental) intervention group,  $P_E$ , can then be obtained using the formula  $P_E = P_C (1 - RRR)$ . Subsequently, the hypothesized  $P_E$  and  $P_C$  may be entered into the formula for the required information size.

Drawing inference about anticipated realistic intervention effects from one intervention area to another may be problematic because an *a priori* estimate may often represent poor approximations of the 'truth'. The clinical trial literature abounds with examples of sample size calculations based on overly optimistic anticipated intervention effects. There is no reason why this should be any different for meta-analysis information size calculations.

If randomised trials have already investigated the effect of an intervention, then a collection of such estimates may be used to better quantify an anticipated intervention effect. However, not all trials provide valid estimates, and caution should be taken to ensure the validity of intervention effects estimates utilised for estimating some anticipated intervention effect.

Many trials yield overestimates of investigated intervention effects due to selective outcome reporting bias and risks of bias (i.e., systematic errors due to inadequate generation of the allocation sequence, inadequate allocation concealment, inadequate blinding, loss to follow-up, or other mechanisms).<sup>13;51-58</sup> Such trials may be classified as trials with *high risk of bias*.<sup>13</sup> Conversely, trials that are likely to yield valid intervention effect estimates may be classified as trials with *low risk of bias*.<sup>13</sup> If evidence on the effect of the investigated intervention is available from a number of trials with low risk of bias, it would be appropriate to base an *a priori* anticipated intervention effect on a meta-analysis of these trials.<sup>6;11;12</sup> However, meta-analytic situations that call for information size calculations will often occur when the evidence is sparse. Even if a number of trials with low risk of bias are available for approximating an anticipated realistic intervention effect, the pooled estimate from these trials may still be subject to considerable random error, time-lag bias, and publication bias. An *a priori* anticipated intervention effect based on the pooled effect estimate from a meta-analysis of trials with low risk of bias is therefore only reliable to the extent that this meta-analysis

can be considered free of large random errors. Furthermore, it is only valid to the extent it can be considered free of time-lag bias and publication bias.

It is not possible to recommend one technique for defining intervention effects for information size calculations. Rather, information size considerations should be based on ranges of plausible control group event proportions, intervention effects, and suitable type I and type II errors. Adequate sample size considerations for a single clinical trial do not just amount to one single number. Instead a range of plausible sample sizes are produced from a range of plausible treatment effects, control group event rates, and type I and type II errors, thus providing a reasonable ballpark interval in which the number of patients need to lie in order to yield a conclusive clinical trial. From produced range of sample sizes, one would select one primary and let the remaining act as sensitivity sample size (power) calculations. We recommend that information size considerations for meta-analysis follow the same construct. Low-bias risk  $P_C$  and RRR estimates could readily be combined with a range of a priori 'realistic' best and worst case intervention effects, thus providing a ballpark interval in which the meta-analysis information needs to lie in order to yield conclusive meta-analytic inferences.

### *Limitations*

The required information size for a meta-analysis (whether determined as the required number of patients, events, or statistical information) comes with a number of limitations. In randomised clinical trials, it is reasonable to assume the distribution of prognostic factors in the randomised patients resembles that of the target population. In systematic reviews with meta-analyses, trials are typically included on the basis of a few inclusion criteria that are decided upon in the protocol stage of the systematic review. Because inclusion (and exclusion) criteria in clinical trials are almost never identical and because trials typically vary in sample sizes, meta-analysts and systematic review authors are unlikely to have control over the distribution of prognostic factors. Even when some systematic review inclusion criteria are altered for an update, authors will not be able to accurately predict the distribution of prognostic factors across newly published trials. Baseline prognostic factors can have a

considerable impact on incidence rates in a control group. In this situation, it may be appropriate to make an *a priori* attempt at quantifying the difference between the baseline incidence in the meta-analysis population and that in the target population, and perform *post hoc* sensitivity analyses if necessary.

Minimally important comparative intervention effects (also known as minimally important differences) may not always be similar across the included trials. For example, if the investigated patient populations across trials experience different risks of adverse events, the minimally important difference may also differ across trials. This variation is the result of clinical intent. For any medical intervention, the chance of benefit needs to outweigh any increased risk of harm. A population with greater risk of harm will need a greater chance of benefit to make a treatment worthwhile. When minimally important differences vary across trials, information size considerations may still be sensible. However, it is important to remember that inference drawn about the conclusiveness of a meta-analysis can only be generalized to the patient population for which the *a priori* minimally important difference apply.

When the required information size is to be defined by the required number of patients or events, the problem of unpredictable heterogeneity may be dealt with by anticipating some appropriate maximum degree of heterogeneity and adjusting the required information size accordingly.<sup>4</sup> The apparent limitation of this approach is that the degree of expected heterogeneity is both difficult to guess and estimate when only a few clinical trials are available. Although we recommend sensitivity analysis on the degree of heterogeneity adjustment, such analyses may still be inappropriate if the anticipated degree(s) of heterogeneity does not reflect the actual degree of heterogeneity which the meta-analyses will incur as more trials are accumulated.

When the required information size is defined by the required statistical information, the formula for the required information size does not require an estimate of the anticipated degree of heterogeneity. Rather, the actual information in the meta-analysis (the estimated statistical information) directly incorporates the heterogeneity through the estimated between-trial variation.

This, however, presents a limitation in that the accumulated statistical information is only reliable to the extent the estimate of the between-trial variance is reliable. Possible solutions to this problem involve the use of more complex methodology to adjust the uncertainty associated with estimating the between-trial variation. One option is to use the random-effects approach by Biggerstaff-Tweedie which incorporates the uncertainty associated with estimating between-trial variance when using the conventional DerSimonian-Laird estimator (see section 2.1.3).<sup>41</sup> Another option is to apply Bayesian meta-analysis, where a prior distribution is elicited for the between-trial variance parameter.

### **2.2.2. The cumulative test statistic (Z-curve)**

As mentioned in section 2.1.2., meta-analysis test for 'statistical significance' uses a Wald-type test statistic. This statistic is given by the log of the pooled intervention effect divided by its standard error,<sup>13</sup> and is commonly referred to as the Z-statistic or the Z-value. Under the assumption that the two investigated interventions do not differ (the *null hypothesis*), the Z-value will approximately follow a standard normal distribution (a normal distribution with mean 0 and standard deviation 1). The larger the absolute value of an observed Z-value, the stronger is the statistical evidence that the two investigated interventions do differ. If the absolute observed Z-value is substantially larger than 0, it is usual to conclude that the observed difference between the effect of the two interventions cannot solely be explained by the play of chance. In this situation, the difference between the two interventions is described as 'statistically significant'. By definition, a P-value is the probability of finding the observed difference, or one more extreme, if the null hypothesis was true. In practice, the P-value is the value that we use to assess statistical significance. The P-value is obtained from the Z-value (see section 2.1.2 for the mathematical details); these two measurements represent two different ways of communicating the same information, and they are inter-changeable. For example, a two-sided P-value smaller than 5% is the same thing as an absolute Z value larger than 1.96, and vice versa.

Every time a meta-analysis is updated, a new Z-value is calculated. A series of consecutive Z-values therefore emanates from a series of meta-analysis updates. To inspect the evolution of significance tests, the series of Z-values can be plotted with respect to the accumulated information (accumulated patients, events, or statistical information), thus producing a curve which is commonly referred to as the *Z-curve*.<sup>1;2;4;6;11;12</sup>

### ***2.2.3. Problems with significance testing in meta-analysis***

As mentioned in chapter 1, conventional significance testing in meta-analysis fails to relate observed test statistics and P-values to the strength of the available evidence and to the number of repeated significance tests.<sup>1-4;6;11;12</sup>

The consequence of this omission is an increased risk of obtaining a false positive meta-analytic result. This section provides basic to intermediate statistical and conceptual descriptions of significance testing in meta-analysis and the problems that result from failing to incorporate the strength of evidence and the number of repeated significance tests into the process.

#### ***General criteria for significance testing***

Conventional significance testing operates with a maximum risk of type I error,  $\alpha$ , which also functions as the threshold for when P-values are considered evidence of statistical significance. P-values and Z-values are interchangeable in the assessment of statistical significant. As mentioned above, for every P-value threshold,  $\alpha$ , there exists a corresponding Z-value threshold,  $Z_{\alpha}$ . For example, if we desire a maximum two-sided type I error risk of 5% we should only consider absolute Z-values larger than 1.96 as evidence of statistical significance. But if we desire a maximum two-sided type I error of 1% we should only consider absolute Z-values larger than 2.58 as evidence of statistical significance.

Let  $\Pr(X|Y)$  denote the probability that the event X occurs given that event Y is true (or has occurred), let  $|Z|$  denote the absolute value of Z. In general, we face the challenge of appropriately determining a threshold,  $c$ , that will make the following equations true

$$Pr(|Z| \geq c \mid H_0 \text{ is true}) \leq \alpha \quad (2)$$

$$Pr(|Z|=c \mid H_0 \text{ is true}) = \alpha \quad (3)$$

For the remaining theoretical sections on repeated significance testing (sections 2.2.2 to 2.2.5), we will assume that all statistical tests are two-sided. We will also assume that all test statistic values,  $Z$ , are absolute values. We assume the latter because the involved algebra becomes much simpler by doing so. For example, in defining two-sided thresholds for a non-absolute test statistic, one would need to consider the probability that  $Pr(Z \leq -c \text{ or } Z \geq c \mid \dots)$  rather than  $Pr(|Z| \geq c \mid \dots)$ .

#### *Problems with repeated significance testing*

Conventional single significance tests can be considered reliable if 'enough' data has accumulated. In meta-analysis, a single significance test can be considered reliable once the required information size is surpassed.<sup>1-4;6;11;12;20;59</sup> If we perform a single test for statistical significance at or after a meta-analysis has surpassed its required information size, statistical significance testing simply entails determining an appropriate threshold,  $c$ , that will make equations (2) and (3) true. For example, for  $\alpha=5\%$  we would consider  $c=1.96$  appropriate if the meta-analysis data had not previously been subjected to significance testing.

When a cumulative meta-analysis is subjected to significance testing more than once (before surpassing its required information size), the situation becomes more complex. Consider the example where a meta-analysis is updated once and where the conventional 5% maximum type I error is used. In this situation, the first meta-analysis yields a  $Z$ -value,  $Z_1$ , and the meta-analysis update yields another,  $Z_2$ . If the first meta-analysis yields a  $Z$ -value larger than 1.96, the two investigated interventions are declared significantly different. However, if the first meta-analysis is not significant (i.e.,  $Z_1 < 1.96$ ), the two interventions can still be declared statistically significant if the meta-analysis update yields a  $Z$ -value larger than 1.96 (i.e., if  $Z_2 \geq 1.96$ ). By the laws

of basic probability theory, the probability that the two interventions will be declared statistically significant under the null hypothesis is:

$$\begin{aligned}\Pr(H_0 \text{ rejected}) &= \Pr(|Z_1| \geq 1.96 \text{ or } |Z_2| \geq 1.96) \\ &= \Pr(|Z_1| \geq 1.96) \cdot \Pr(|Z_2| \geq 1.96 \mid |Z_1| < 1.96)\end{aligned}$$

It can be shown that this expression is always larger than the desired 5% (see appendix A.3.1). In general, repeated significance testing using single test thresholds will always lead to an exaggeration of the type I error, and the larger the number of (repeated) significance tests employed on accumulating data, the worse the exaggeration of the type I error becomes.<sup>30</sup> For meta-analysis data, simulation studies have demonstrated that repeated significance testing result in a type I error of 10% to 30% when the conventional  $\alpha=5\%$  threshold, 1.96, is used to test for statistical significance at every update.<sup>7;8;10;31</sup>

#### **2.2.4. The $\alpha$ -spending function and trial sequential monitoring boundaries**

One solution to the problem outlined in section 2.2.3. is to adjust the thresholds for the Z-values, allowing the type I error risk to be restored to the desired maximum risk.<sup>1;2;6;17</sup> In the two tests example, we would thus need to find two thresholds,  $c_1$  and  $c_2$ , for which

$$\Pr(|Z_1| \geq c_1 \text{ or } |Z_2| \geq c_2) \leq \alpha$$

is satisfied under the null hypothesis. This is equivalent to finding two maximum type I error risks,  $\alpha_1$  and  $\alpha_2$ , that sum to  $\alpha$  and where

$$\Pr(|Z_1| \geq c_1) \leq \alpha_1$$

$$\Pr(|Z_2| \geq c_2 \mid |Z_1| < c_1) \leq \alpha_2$$



under the null hypothesis. In the general situation where repeated significance testing is employed  $k$  times (i.e., where one initial meta-analysis and  $k-1$  updates are performed), we would need to find thresholds  $c_1, \dots, c_k$  for each of the  $k$  significance tests that will ensure

$$\Pr(|Z_1| \geq c_1 \text{ or } |Z_2| \geq c_2 \text{ or } \dots \text{ or } |Z_k| \geq c_k) \leq \alpha$$

under the null hypothesis. This is equivalent to finding  $k$  maximum type I error risks,  $\alpha_1, \dots, \alpha_k$ , that sum to  $\alpha$  and where

$$\begin{aligned} \Pr(|Z_1| \geq c_1) &\leq \alpha_1 \\ \Pr(|Z_2| \geq c_2 \mid |Z_1| < c_1) &\leq \alpha_2 \\ \Pr(|Z_3| \geq c_3 \mid |Z_1| < c_1 \text{ and } |Z_2| < c_2) &\leq \alpha_3 \\ &\vdots \\ \Pr(|Z_k| \geq c_k \mid |Z_1| < c_1 \text{ and } \dots \text{ and } |Z_{k-1}| < c_{k-1}) &\leq \alpha_k \end{aligned}$$

under the null hypothesis.

The collation of thresholds for the Z-curve is referred to as *monitoring boundaries*, or *group sequential monitoring boundaries* (a series of boundaries applied to sequence of tests on cumulative *groups* of patients randomised in a clinical trial).<sup>17;30;44</sup> In meta-analysis, such boundaries are applied to a sequence of trials, and we therefore refer to them as *trial sequential monitoring boundaries*.<sup>6</sup> The combination of meta-analysis and trial sequential monitoring boundaries is referred to as *trial sequential analysis*.<sup>6</sup>

Trial sequential monitoring boundaries require pre-specification of the  $k$  maximum type I error risks,  $\alpha_1, \dots, \alpha_k$ , as well as intensive numerical integration for their application.<sup>60</sup> One simple method for assigning values for the  $\alpha_1, \dots, \alpha_k$  type I error risks is the  *$\alpha$ -spending method* (or  *$\alpha$ -spending function*).<sup>1;2;17;30</sup> This method is implemented in the TSA program. The  *$\alpha$ -spending function* is a monotonically increasing function of time that can be used for appropriately assigning maximum type I error risks  $\alpha_1, \dots, \alpha_k$  at each

significance test according to the amount of information accumulated.<sup>16;17</sup> The independent variable is defined by the information fraction (IF); this is calculated by dividing the accumulated information by the required information size (e.g., the accumulated number of patients divided by the required number of patients).<sup>6;15;17</sup> The dependent variable (the function) is the cumulative type 1 error; this gives the amount of error that should be considered the maximum when defining significance at the given IF. As IF increases – i.e., as the amount of accumulated information increases – the size of ‘acceptable’ type 1 error also increases. The function provides a way to quantify the risk of random error allowed at any given IF, in order to ensure that the overall risk of random error – after the IS has been reached – stays below 5%. The monotonically increasing function corresponds to a monotonically decreasing threshold for statistical significance measured by the test statistic Z.

The  $\alpha$ -spending function is defined from 0 to 1 (0 being the point where 0 patients have been randomised, and 1 being the point where the accumulated information equals the required information size).<sup>16;17</sup> The  $\alpha$ -spending function of 0 is always equal to 0:  $\alpha(0)=0$ ; at this point, no information has been accumulated. The  $\alpha$ -spending function of 1 is always equal to  $\alpha$ :  $\alpha(1)=\alpha$ ; at this point, all of the required information has been accumulated and the total amount of alpha error is whatever was defined as total acceptable type 1 error overall (usually 5%). At any point between 0 and 1 (for the information fraction at the time of a significance test  $i$  ( $IF_i$ )) the  $\alpha$ -spending function is equal to the total maximum type I error risk that has arisen from the thresholds chosen for all significance tests until and including the  $i$ -th significance test. In other words, the  $\alpha$ -spending function is equal to how much type 1 error has been ‘spent’. In notation:  $\alpha(IF_i)=\alpha_1+\alpha_2+\dots+\alpha_i$ , and thus

$$\begin{aligned} \Pr(|Z_1| \geq c_1) &\leq \alpha_1 = \alpha(IF_1) \\ \Pr(|Z_2| \geq c_2 \mid |Z_1| < c_1) &\leq \alpha_2 = \alpha(IF_2) - \alpha(IF_1) \\ \Pr(|Z_3| \geq c_3 \mid |Z_1| < c_1 \text{ and } |Z_2| < c_2) &\leq \alpha_3 = \alpha(IF_3) - \alpha(IF_2) \\ &\vdots \\ \Pr(|Z_k| \geq c_k \mid |Z_1| < c_1 \text{ and } \dots \text{ and } |Z_{k-1}| < c_{k-1}) &\leq \alpha_k = \alpha(IF_k) - \alpha(IF_{k-1}) \end{aligned}$$

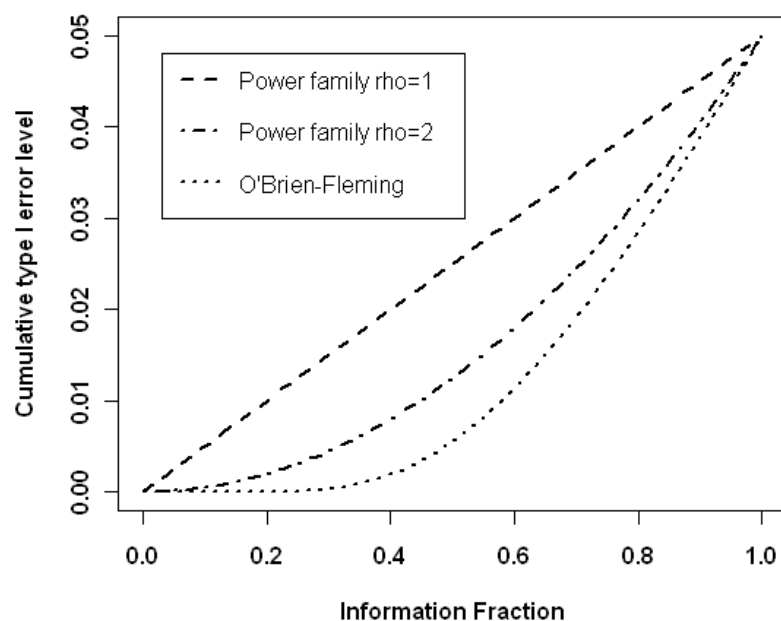
The actual  $\alpha$ -spending function used can be any monotonically increasing function.<sup>16;17</sup> One well-known example is  $\alpha(t)=t\cdot\alpha$ .<sup>16;17;30</sup> When all significance tests are performed at an equal distance (with respect to the information fraction scale), this  $\alpha$ -spending function will yield equal thresholds for the Z-values (i.e.,  $c_1=c_2= \dots=c_k$ ). This adjustment was first proposed by Pocock. A more general  $\alpha$ -spending approach is the *power family*  $\alpha$ -spending function defined as  $\alpha(t)=t^\rho\cdot\alpha$ .<sup>16;17;30</sup> Power family  $\alpha$ -spending functions, where  $\rho>1$  and where all significance tests are performed at equal distance, will yield more conservative thresholds for early significance tests than for later significance tests. In general, the thresholds for (absolute values of) the Z-curve will be monotonically decreasing when the  $\alpha$ -spending function is convex and all significance tests are performed at equal distance.<sup>16;17;30</sup> Monotonically decreasing thresholds (which result from the monotonically increasing functions) are desirable because the impact of random error is typically inversely proportional to the amount of accumulated information. Although an infinite combination of decreasing thresholds exists, some sets of thresholds may be preferable.

From advanced probability theory, the  $\alpha$ -spending function that yield theoretically optimal thresholds is given by the expression

$$\alpha(IF)=2-2\Phi\left(Z_{\alpha/2}/\sqrt{IF}\right)$$

where  $\Phi$  is the standard normal cumulative distribution function.<sup>16;17;30</sup> The type of boundaries produced by this  $\alpha$ -spending function were first proposed for equal increments of IF by O'Brien and Fleming.<sup>61</sup> Lan and DeMets later proposed the above  $\alpha$ -spending function to allow for flexible increments in IF.<sup>16;17;30</sup> For this reason, the above  $\alpha$ -spending function is typically referred to as the Lan-DeMets implementation of the O'Brien-Fleming  $\alpha$ -spending function. Often, the monitoring boundaries produced by this alpha spending function are simply referred to as the Lan-DeMets monitoring boundaries or the O'Brien-Fleming monitoring boundaries. For the remainder of this manual, we will refer to them as O'Brien-Fleming monitoring boundaries. Currently, the

O'Brien-Fleming  $\alpha$ -spending function is the only  $\alpha$ -spending function implemented in the TSA software.



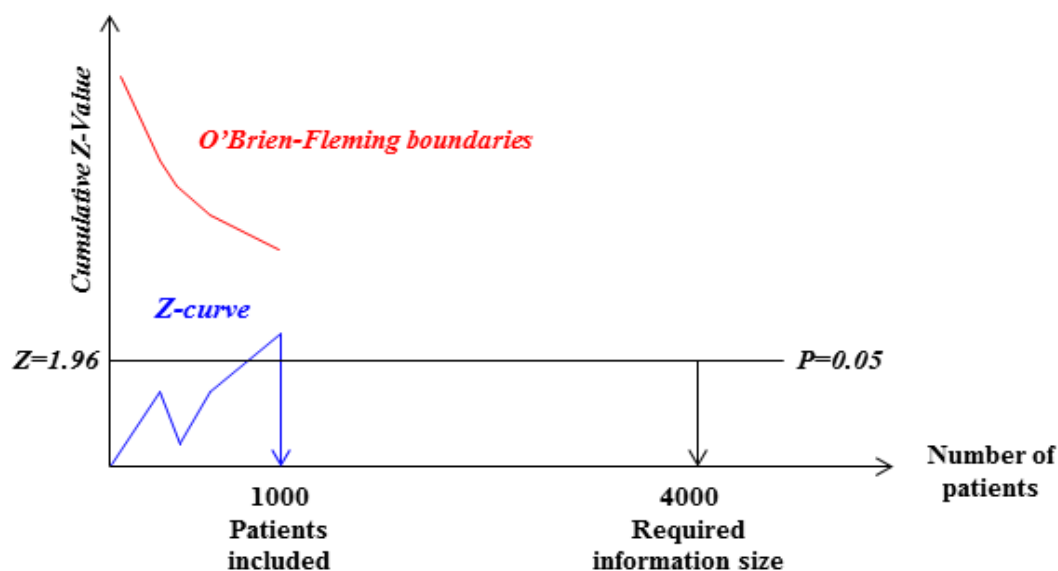
**Figure 6** The shape of the power family  $\alpha$ -spending functions with  $\rho=1$  and  $\rho=2$  and the O'Brien-Fleming  $\alpha$ -spending function.

As shown in figure 6, the O'Brien-Fleming  $\alpha$ -spending function is an exponentially increasing function. It produces conservative boundaries at early stages where only limited amount of data has been accumulated, and more lenient boundaries as more data are accumulated.

The O'Brien-Fleming boundaries have been recommended by methodological experts as the preferred choice in most randomised clinical trials where repeated significance testing on accumulating data is performed.<sup>30,62</sup> In meta-analysis, where the risk of random error (and time-trend biases) is of particular concern at early stages (i.e., in meta-analyses including a small number of patients and events), the O'Brien-Fleming boundaries have been the preferred choice as well.<sup>1;2;4;6;11;12</sup>

There are two reasons for this preference. First, if the heterogeneity adjustment of the required information size is based on a reasonable *a priori* estimate of the anticipated degree of heterogeneity, the O'Brien-Fleming

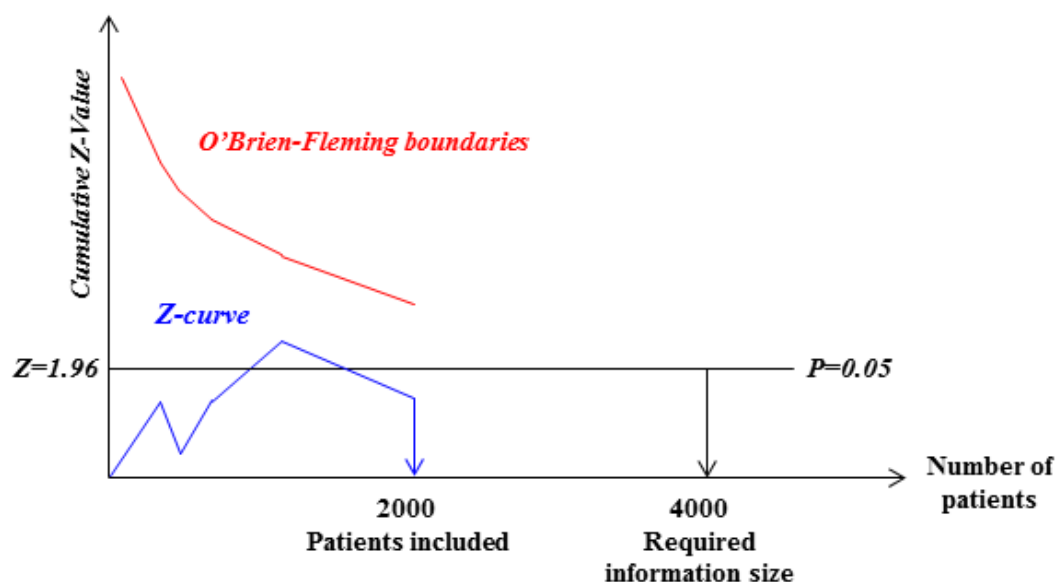
boundaries will naturally account for the degree of fluctuations that the meta-analytic inferences will incur due to random error and heterogeneity. Second, as long as subsequent significance tests are performed at a reasonable distance on the information axis (e.g., at least 1% of the required information size apart), the O'Brien-Fleming boundaries remain relatively unaffected by the number of previous significance tests. This second property is desirable in the setting of meta-analysis because it is not always clear how often a meta-analysis has been subjected to significance testing as a result of updating. For example, some meta-analyses may include different but highly overlapping data because the inclusion criteria have been modified in connection with updates of a systematic review. Other monitoring boundaries, such as a set of the monitoring boundaries based on the power family alpha spending function with  $\rho=2$ , could yield discrepant inferences about statistical significance if, for example, the monitoring boundaries accounted for 2 previous updates as opposed to 4.



**Figure 7** Example of an inconclusive meta-analysis after four cumulative meta-analyses.

Figure 7 shows an example of the use of the O'Brien-Fleming boundaries. In this meta-analysis, the required information size is 4000 patients, but the obtained information is only 1000 patients. The final Z-value is larger than 1.96. Using the conventional single test threshold, this Z-value would have led to a conclusion of statistical significance. Using the O'Brien-Fleming boundaries, a greater value of Z is required – at this information size – in

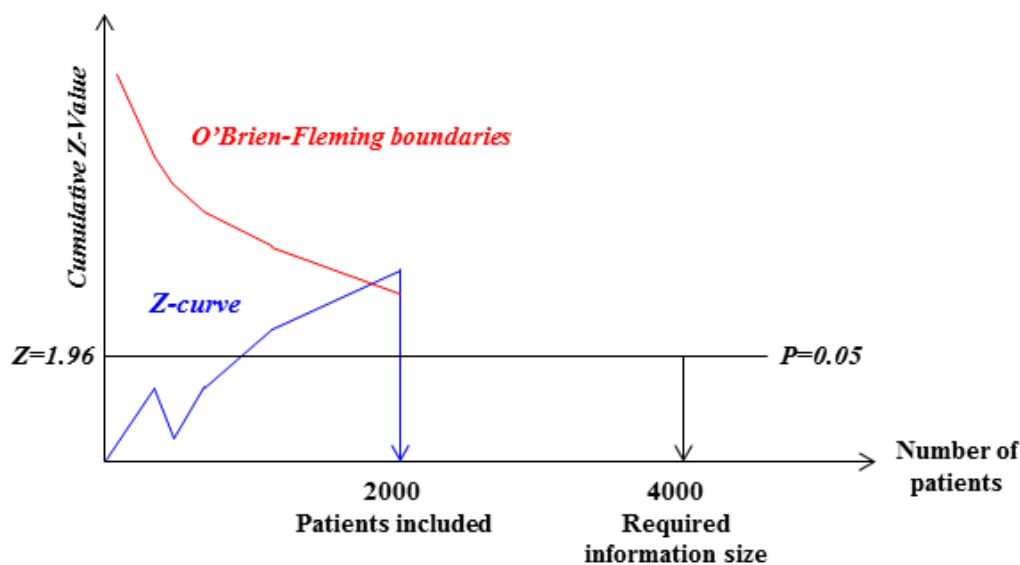
order to conclude statistical significance. The boundaries are not crossed and the meta-analysis is therefore inconclusive.



**Figure 8** Example of a meta-analysis including a false positive Z-value at the fifth cumulative significance testing.

In the example given in figure 8, the required information size is again 4000 patients and the obtained information is now 2000 patients. The final Z-value is smaller than 1.96; this result would have been inconclusive using either conventional or boundary techniques. There are, however, preceding Z values that had been calculated in the cumulative process, including one with a value greater than 1.96. This example illustrates how a cumulative Z curve could cross the conventional threshold for significance in an early meta-analysis, only to be declared not significant in a later meta-analysis. O'Brien-Fleming boundaries can prevent such premature false positive conclusions.

In the example given in figure 9, the required information size and the attained information size are the same as those in figure 8. Here, the Z-value calculated at the fifth significance test is 'extreme enough'; the Z-curve crosses the O'Brien-Fleming boundaries, and the meta-analysis can be declared as conclusive with regard to the anticipated intervention effect leading to the required information size.



**Figure 9** Example of a meta-analysis that becomes conclusive according to the O'Brien-Fleming boundaries after the fifth cumulative significance testing.

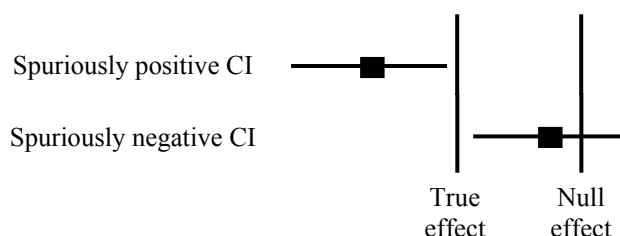
In the above examples (figure 7-9), the monitoring boundaries are constructed only for the positive half of the y-axis. Two-sided symmetrical significance testing boundaries can be constructed on both the negative and positive half of the y-axis. The TSA program allows for both one and two-sided significance testing. When the outcome measure for binary data meta-analysis is defined as a failure (e.g., death or relapse), Z-values on the upper half of the y-axis will indicate benefit of the experimental intervention, whereas Z-values on the lower half will indicate harm.

The monitoring boundaries' values for the Z-curve are a function of the alpha spending function; they are calculated by numerical recursive integration according to Reboussin et al.<sup>60</sup> Though all boundary values are discrete points calculated for each cumulative update of the meta-analysis, the TSA program connects these points and creates one continuous boundary line for better visual interpretation.

### **2.2.5. Adjusted confidence intervals following trial sequential analysis**

Just as repeated significance tests affects the overall type I error, it also affects the construction of confidence intervals. For example, when we assume that our pooled estimate of effect is normally distributed (as we typically do in meta-analysis), we form a 'naïve' symmetric 95% confidence

interval  $\hat{\mu} \pm 1.96 \cdot se(\hat{\mu})$ , where  $\hat{\mu}$  denotes our estimated meta-analysed intervention effect and  $se(\hat{\mu})$  denotes its associated standard error. However, if a meta-analysis is subjected to repeated statistical evaluation, and thus, produces a series of confidence intervals over time, the probability that all of these confidence intervals will contain the ‘true’ overall effect is certainly less than 95%. That is, if we construct a series of naïve symmetric  $(1-\alpha)\%$  confidence intervals,  $\hat{\mu} \pm z_{1-\alpha/2} \cdot se(\hat{\mu})$ , the probability that all these confidence intervals will contain the ‘true’ overall effect is certainly less than  $(1-\alpha)\%$ . Thus, when a meta-analysis is subjected to repeated statistical evaluation, there is an exaggerated risk that the ‘naïve’ confidence intervals will yield spurious inferences. When some underlying ‘true’ intervention effect exists, spurious inferences based on confidence intervals can occur as either of the two scenarios illustrated in figure 10.



**Figure 10** Example of spuriously positive and spuriously negative confidence interval inferences.

When there is no intervention effect, the confidence intervals will yield spurious inferences if they preclude the null effect. This situation is identical to a false positive significance test (see section 2.2.4).

Similar to adjustment for repeated significance testing, the confidence intervals can be adjusted according to the strength of the available information (e.g., the number of patients) and the number of statistical evaluations. If we let  $l$  and  $u$  denote the lower and upper limit of some naïve confidence interval with coverage  $1-\alpha$ , we know that

$$\Pr(l \leq \mu \leq u) = 1-\alpha$$



When a meta-analysis is subjected to repeated statistical evaluation, the repeated naïve confidence intervals will not yield the desired coverage. Thus, we need to establish a series of intervals that will achieve the desired coverage. Assume that a meta-analysis is subjected to statistical evaluation  $k$  times up till the point where it surpasses its required information size. Let  $l_1, l_2, \dots, l_k$  and  $u_1, u_2, \dots, u_k$  denote the lower and upper confidence interval limits for each of the  $k$  times the meta-analysis was subjected to statistical evaluation. To maintain the desired coverage, these limits would have to satisfy:

$$\Pr(l_1 \leq \mu \leq u_1, l_2 \leq \mu \leq u_2, \dots, l_k \leq \mu \leq u_k) \geq 1 - \alpha$$

And thus, any single one of these  $k$  intervals, say  $j$ , would have to satisfy:

$$\Pr(l_j \leq \mu \leq u_j) \geq 1 - \alpha$$

It is clear from the above that the  $\alpha$ -level for each repeated confidence interval cannot exceed the overall maximum  $\alpha$ . Further, the respective  $\alpha$ -levels for each of the repeated confidence intervals should sum to the overall maximum  $\alpha$ . Thus, by controlling the overall  $\alpha$ -level, we can control the overall coverage. The framework for controlling the overall  $\alpha$ -level has already been developed in the previous section (2.2.4), and is easily applied to repeated confidence intervals. Naïve confidence intervals are obtained using the formula  $\hat{\mu} \pm z_{1-\alpha/2} \cdot se(\hat{\mu})$  because we know that  $z_{\alpha/2} \leq \hat{\mu} / se(\hat{\mu}) \leq z_{1-\alpha/2}$  with approximately  $(1-\alpha)\%$  probability (under the null hypothesis), and hence:

$$z_{\alpha/2} \leq Z \leq z_{1-\alpha/2},$$

where  $Z$  denotes the Z-value for the statistical significance test. By replacing  $z_{\alpha/2}$  and  $z_{1-\alpha/2}$  by the thresholds that constitute the statistical monitoring boundaries,  $c_1, c_2, \dots, c_k$ , and isolating for  $\hat{\mu}$ , we have constructed a simple expression for repeated confidence intervals which will maintain good control of the coverage. For any single one of the  $k$  confidence intervals, say  $j$ , the expression for the confidence interval is:

$$\hat{\mu} \pm c_j \cdot se(\hat{\mu})$$

And we have

$$\Pr(\hat{\mu} - c_1 \cdot se(\hat{\mu}) \leq \mu \leq \hat{\mu} + c_1 \cdot se(\hat{\mu}), \dots, \hat{\mu} - c_k \cdot se(\hat{\mu}) \leq \mu \leq \hat{\mu} + c_k \cdot se(\hat{\mu})) \geq 1 - \alpha$$

All of the above easily generalises to one-sided confidence intervals.

The TSA software provides the option of calculating the confidence interval for the last of a series of statistical evaluations (see chapter 4).

### **2.2.6. The law of the iterated logarithm**

Another solution to the problem of repeated significance testing outlined in section 2.2.3. is to penalise the Z values according to the strength of the available evidence and the number of repeated significance tests.<sup>7;8</sup> In advanced probability, there exists a theorem, *the law of the iterated logarithms*, which tells us that if we take a standard normally distributed variable, such as a Z-value, and divide it by the logarithm of the logarithm of the number of observations in the data, there will be a 100% probability that this fraction will assume a value between  $-\sqrt{2}$  and  $\sqrt{2}$ . In the context of statistical testing, this law can be utilised to control exaggeration of type 1 error in meta-analysis due to repeated significance testing. Dividing a standard normally distributed test statistic by the logarithm of the logarithm of the information available, provided enough data has accumulated, can provide good control of the 'behaviour' of the employed statistical test. Lan et al. applied this theory, introducing a penalty for the Z-values obtained at each significance test and creating adjusted (penalised) Z-values,  $Z^*$ , given by

$$Z_j^* = \frac{Z_j}{\sqrt{\lambda \ln(\ln(I_j))}}$$

where  $Z_j$  is the conventional Z-value,  $I_j$  is the cumulative statistical information at the  $j$ -th significance test (see section 2.2.1. under *alternatives to*

*accumulating number of patients*), and  $\lambda$  is some constant that will ensure good control of the maximum type I error.<sup>8</sup> Lan et al. used simulation to estimate proper choices of the constant,  $\lambda$ , for continuous data meta-analysis,<sup>8</sup> and Hu et al. did the same for dichotomous data meta-analysis.<sup>7</sup> For continuous data meta-analysis, Lan et al. found that  $\lambda=2$  would generally exhibit good control of the type I error, when using a desired maximum type I error of  $\alpha=5\%$  for a two-sided statistical test (i.e.,  $\alpha=2.5\%$  for each side).<sup>8</sup> That is, when  $Z^*$  was evaluated based on the conventional criteria for statistical significance (i.e.,  $|Z^*| \geq 1.96$  means statistical significance at two-sided  $\alpha=5\%$ ). For dichotomous data meta-analysis, Hu et al. estimated appropriate choices of  $\lambda$  for different maximum type I error levels and different effect measures.<sup>7</sup> Their simulation results lead to the recommended  $\lambda$  values presented in table 2.

Effect measure	Max. type I error (corresponding threshold)		
	$\alpha=0.01$ (c=2.33)	$\alpha=0.025$ (c=1.96)	$\alpha=0.05$ (c=1.65)
Risk difference	$\lambda=3$	$\lambda=1.5$	$\lambda=1.5$
Relative risk	$\lambda=3.5$	$\lambda=2$	$\lambda=2$
Odds ratio	$\lambda=3.5$	$\lambda=2$	$\lambda=2$

**Table 2** Recommended  $\lambda$  values for penalising Z values with the law of the iterated logarithm

These  $\lambda$  values pertain only to the ranges of study sizes, control group event proportion, and between-trial variation used in the simulations, and may therefore not be applicable to all meta-analysis scenarios.<sup>7;8</sup> For example, the minimum event proportion in the control groups used in the simulations was 0.05. Many important clinical conditions yield control group event proportions lower than 0.05. In addition, none of the simulations incorporated time trend bias such as time lag bias and publication bias. Such biases have a considerable impact on significance tests in meta-analyses. Further, as previously noted (section 2.2.1 - Limitations), statistical information relies on accurate and reliable estimation of the between-trial variance. If the between-

trial variance is underestimated (for example due to time-lag bias), the penalised Z-statistic will be artificially large. For the above reasons, it is reasonable to assume that the recommended  $\lambda$  values in table 2 constitute the very minimum of a range of appropriate choices. Appropriate  $\lambda$  values for dichotomous data meta-analyses including only a small number of trials, patients and/or events are probably higher than those recommended by Hu et al.

### **2.2.7. The $\beta$ -spending function and futility boundaries**

When a result in a meta-analysis is found to be non-significant, it is important to assess whether this non-significance is due to lack of power or whether it is due to underlying equivalency between the interventions.

The statistical exercise of testing for equivalency – i.e., testing for both non-superiority and non-inferiority of a given intervention – is commonly referred to as *futility testing*.<sup>30</sup> The statistical test thresholds that arise from this exercise are referred to as *futility boundaries*. When a Z-curve crosses the futility boundaries, we can accept that the two interventions do not differ more than the anticipated intervention effect.

Meta-analyses that have already surpassed their required information should have enough power to demonstrate superiority of one intervention over the other. For this sub-section, we will consider only non-significant meta-analyses that have not surpassed their required information size. Further, we no longer consider all Z values as absolute. Instead we make the distinction of positive Z values indicating that the experimental intervention is superior to the control intervention and negative Z values indicating that the experimental intervention is inferior to the control intervention. The following section deals first with non-superiority testing, followed by non-inferiority testing and futility testing in general.

At any point, a meta-analysis may yield a Z value that is not statistically significant in favour of the experimental intervention. However, only when this Z value lies 'sufficiently below' the threshold for statistical significance (in

favour of the experimental intervention) can we be confident that the experimental intervention is not superior to the control. To make sense of the above, we must first define what we mean by superior and 'sufficiently below'.

Within the framework of repeated statistical testing, the definition of superiority is linked to the underlying assumption made for the required information size. When calculating the required information size, we assume, *a priori*, an intervention effect,  $\delta$ . The magnitude of this effect represents what we believe to be a *minimally important difference* between the two interventions. Ideally, the size of  $\delta$  should be defined such that anything smaller would be considered clinically, or practically, unimportant and therefore not worth investigating. The value of  $\delta$  depends on the context of the study. For example, a RRR of 10% would usually be considered important if the outcome is mortality, but it may not be considered important if the outcome is nausea.

Before we define what is meant by 'sufficiently below' in the context of repeated statistical testing, consider first the situation where the information contained in a meta-analysis equals its required information size and where statistical testing is performed for the first time. First, let  $H_\delta$  denote the hypothesis that the effect is equal to  $\delta$  - this is the *alternative hypothesis* (in contrast to the *null hypothesis*). Under the assumption that  $H_\delta$  is true, the probability that the meta-analysis will be statistically significant (with the chosen  $\alpha$ -level) is equal to the chosen power,  $1-\beta$ . When the information size has been reached, the probability that the meta-analysis will be falsely negative is equal to  $\beta$ . In this situation, our threshold for statistical significance,  $c$ , which satisfies that:

$$Pr(|Z| \geq c \mid H_0 \text{ is true}) \leq \alpha$$

implicitly becomes our threshold for non-superiority because  $c$  also satisfies:

$$Pr(Z < c \mid H_\delta \text{ is true}) \leq \beta.$$

When repeated statistical testing occurs before a meta-analysis surpasses its required information size, it is also possible to test for non-superiority. This testing can be done by defining thresholds that, under the alternative hypothesis, do not result in an inflation of the total risk of type II error. For example, if we test for non-superiority two times, we need to find thresholds,  $c_1$  and  $c_2$ , for the emerging two subsequent Z values,  $Z_1$  and  $Z_2$ ,

$$\Pr(Z_1 < c_1 \text{ or } Z_2 < c_2) \leq \beta$$

In this situation,  $Z_1$  values smaller than  $c_1$  and  $Z_2$  values smaller than  $c_2$  will be considered 'sufficiently below' the threshold for statistical significance to justify the conclusion of non-superiority. In a more general context, where we might test for non-superiority  $k$  times, we would need to find thresholds  $c_1, \dots, c_k$  which will satisfy

$$\Pr(Z_1 < c_1 \text{ or } Z_2 < c_2 \text{ or } \dots \text{ or } Z_k < c_k) \leq \beta$$

under the alternative hypothesis,  $H_\delta$ . This is equivalent to finding  $k$  maximum type II error risks,  $\beta_1, \dots, \beta_k$ , that sum to  $\beta$  and where

$$\begin{aligned} \Pr(Z_1 < c_1) &\leq \beta_1 \\ \Pr(Z_2 < c_2 \mid Z_1 \geq c_1) &\leq \beta_2 \\ \Pr(Z_3 < c_3 \mid Z_1 \geq c_1 \text{ and } Z_2 \geq c_2) &\leq \beta_3 \\ &\vdots \\ \Pr(Z_k < c_k \mid Z_1 \geq c_1 \text{ and } \dots \text{ and } Z_{k-1} \geq c_{k-1}) &\leq \beta_k \end{aligned}$$

under the alternative hypothesis.

This desire to control the type II error in the context of repeated testing is analogous to the desire to control the type I error. Multiple testing increases the actual amount of error and we need to find a technique to control this increase. Just as it is caused by the same phenomenon, the problem of an increased type II error can be managed using a similar solution. In section

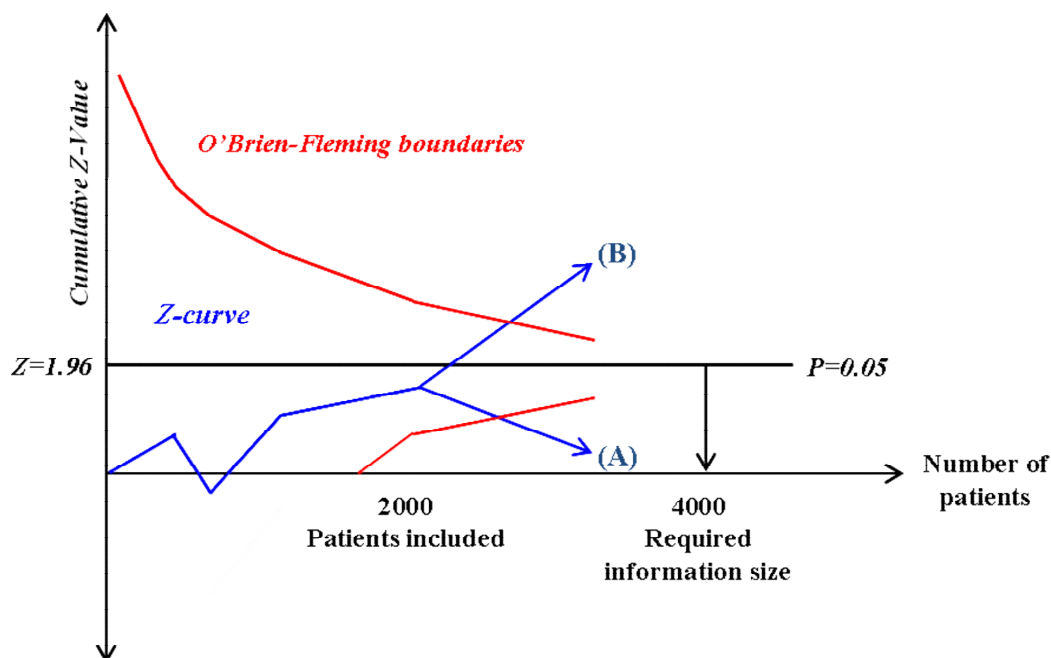
2.2.3, the alpha spending function was described as a technique which can be used to create reasonable boundaries for significance testing. Similarly, the problem of finding repeated non-superiority testing thresholds, which will ensure good control of the type II error, can be solved by introducing the  $\beta$ -spending function. The  $\beta$ -spending function is a monotonically increasing function of time which is used to appropriately assign maximum type II error risks  $\beta_1, \dots, \beta_k$  at each non-superiority test according to the amount of information accumulated. The  $\beta$ -spending function is a function of the information fraction, IF (the accumulated information divided by the required information size), and it is only defined from 0 to 1. The  $\beta$ -spending function of 0 is always equal to 0:  $\beta(0)=0$ , and the  $\beta$ -spending function of 1 is always equal to  $\beta$ :  $\beta(1)=\beta$ . At any point between 0 and 1, the  $\beta$ -spending function is equal to the total maximum type II error risk that has arisen from the thresholds chosen for all non-superiority tests until and including the  $i$ -th test. In other words, the  $\beta$ -spending function is equal to how much type II error has been 'spent'. In notation:  $\beta(IF_i)=\beta_1+\beta_2+\dots+\beta_i$ .

For the same reasons described in section 2.2.4, the O'Brien-Fleming function may also constitute the optimal choice for the beta-spending function. In TSA v.0.8, the only available  $\beta$ -spending function is the O'Brien-Fleming spending function.

Figure 11 shows an example of a meta-analysis including both repeated non-superiority and significance testing. In this meta-analysis, the required information size is 4000 patients. At 2000 patients, the meta-analysis is inconclusive because it has not yet crossed the (upper) boundary for statistical significance or the (lower) boundaries for non-superiority. The dashed extensions of the Z curve illustrate examples of how the meta-analysis could become conclusive at 3000 patients.

In example (A), the Z-curve crosses the non-superiority boundaries (the lower boundaries), in which case, it would be inferred that the experimental intervention is not superior to the control intervention. In example (B), the Z-curve crosses the O'Brien-Fleming significance boundaries for superiority, in

which case, it would be inferred that the experimental intervention is superior to the control intervention.



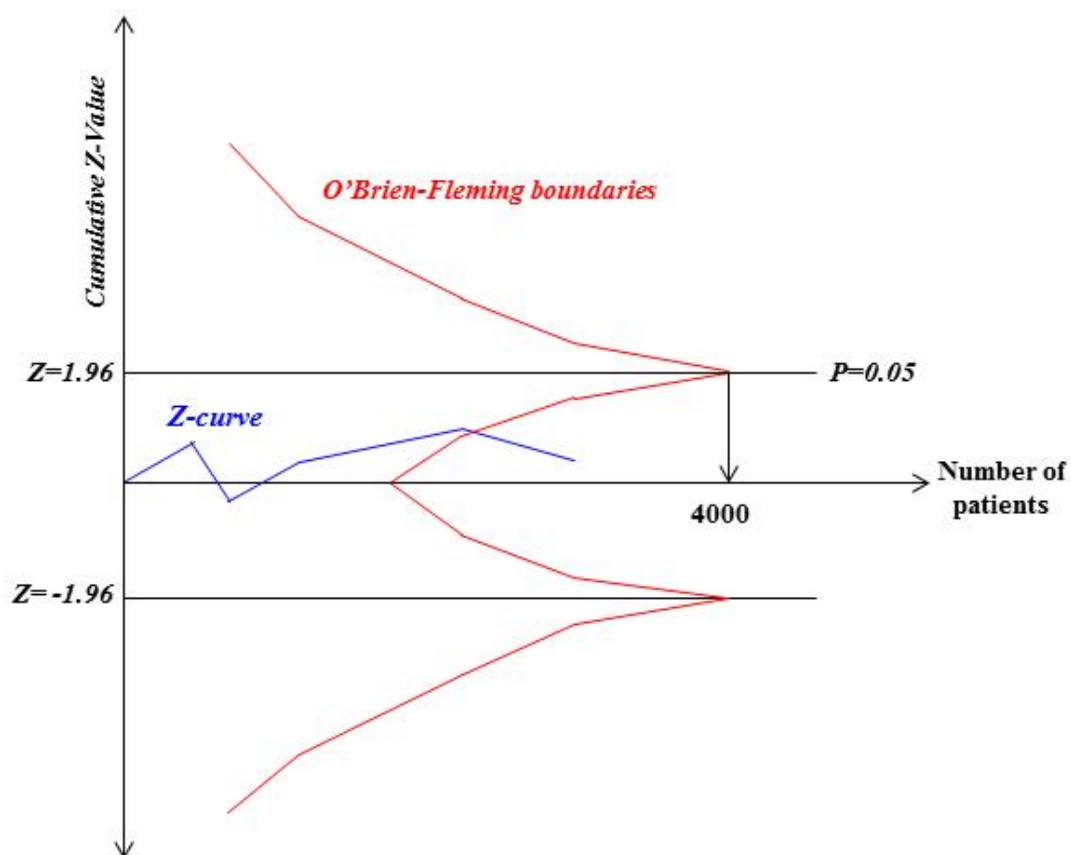
**Figure 11** Example of a meta-analysis including repeated non-superiority (red line) and significance (brown line) testing. The cumulative Z-curve for the first four trials reaches half of the required information size. Two new trials are added to the meta-analysis – (A) showing no effect (and the cumulative Z score now reaches futility) and (B) showing significant benefit of the intervention (and the cumulative Z-score now reaches significance by crossing both the conventional boundary as well as the O'Brian-Fleming boundaries).

Non-superiority boundaries need to be used in conjunction with non-inferiority boundaries in order to assess for equivalence between two groups. Imagine a meta-analysis comparing two groups: group A and group B. If a cumulative Z value falls below the non-superiority threshold, then group A is not better than group B. But it may be worse. If the same cumulative Z value also falls above the non-inferiority threshold, then group A is not worse than group B. In this situation, it can be concluded that group A and B are equivalent. Graphically, this 'area of equivalence' is the area within the two boundaries after they cross – also called the *inner wedge* (see figure 12).

Figure 12 shows an example of a meta-analysis that includes all of the components of TSA that have been discussed: the required information size, two-sided significance testing boundaries, non-superiority futility boundaries and non-inferiority futility boundaries. In this example, the required information size is 4000. At approximately 3000 patients, the Z value falls within the inner



wedge and a conclusion can be made: the intervention effect is not greater than the one anticipated.



**Figure 12** Example of a meta-analysis with repeated non-superiority, non-inferiority and significance testing boundaries.

## **3. Installation and starting the TSA program**

### **3.1. Prerequisites**

The Trial Sequential Analysis (TSA) software is a Java™ program and will therefore run on any operating system that supports Java™ (Microsoft Windows, Mac OS, UNIX, Linux, etc.). The TSA software requires that you have the latest (or at least a recent) version of the *Java Runtime Environment* (JRE) installed on your computer. You can download the JRE for free at [www.java.com](http://www.java.com).

At the time of writing (August 2011), the latest JRE version is 1.6. The TSA software runs well with this version.

### **3.2. Installation**

The TSA software is delivered in a *ZIP archive*. Use any archive tool, such as WinRAR or GZIP, to unpack the archive. In the archive you will find three files named *TSA.jar*, *RM5Converter.jar*, and *TEMPLATES.TPL* along with two folders named *lib* and *samples*.

*TSA.jar* is a Java archive containing the Trial Sequential Analysis application. *RM5Converter.jar* is a Java archive containing an application for converting trial data (presently, however, for dichotomous outcomes only) exported from *Review Manager v.5* into the appropriate data format for TSA. *TEMPLATES.TPL* contains monitoring boundary templates that you can use when you are performing trial sequential analysis on your meta-analysis data. The content of the templates file is controlled through the TSA program. The folder 'lib' contains various external packages used by the TSA program. The folder 'samples' contains '.TSA' files for the examples provided in this manual (see chapter 5).

To install the program, unpack the entire ZIP archive into a folder of your choice on your hard drive. No further steps are required.

### **3.3. Starting TSA**

To start the TSA software, double-click the *TSA.jar* file.

Alternatively the TSA software can be started in a *prompt*. To start the TSA software in a prompt, first start a prompt, browse to the folder in which you have unpacked the TSA software, and type 'java -jar TSA.jar'.

If you are using the Microsoft Windows operating system, you can open a *dos prompt* by first clicking on the 'Start' button (typically lower left corner of the screen), then clicking on 'Run...'. When the 'Run' window pops up, type in 'cmd' (no quotes) in the text field and press OK. The *dos prompt* should appear. Use the *cd* (change directory) command to browse to the folder in which you have unpacked the TSA software. For example, if you created a folder named *TSA* within the *Program Files* folder on your C drive and unpacked the TSA software to this folder, you should first change the directory to the *TSA* folder in the *dos prompt*. This can be done by typing 'cd C:\Program Files\TSA' (no quotes). After the directory in the *dos prompt* has been changed, type 'java -jar TSA.jar'.

### **3.3.1. Why doesn't TSA start?**

If you are having trouble starting the TSA software, there are several possible reasons for this. Below is a check list to help identify the most likely reasons:

is the JRE installed on your system?

is the installed JRE version at least 1.6?

did you extract all the files from the ZIP archive?

did you rename, move, or delete any of the unpacked files or folders?

If a different program (other than TSA) starts when double-clicking the TSA.jar file, this means that the .jar file name extension is not associated to Java (JRE). If this happens, you can either try to start the program manually using a prompt (see above), or you can try to change the file name association. If you are using Windows, you can change the association by following the steps below:

- open an Explorer window (e.g., double-click on My Computer) and click the 'Tools' menu
- select 'Folder Options...' and go to the 'File Types' tab
- find the JAR extension in the list
- click 'Change'
- select 'Java(TM) Platform SE binary' from the list and click OK
- If 'Java(TM) Platform SE binary' is not in the list, click 'Browse' and locate the javaw.exe in the JRE's bin folder. Its default path is: C:\Program Files\Java\jre6\bin.

If your operating system is not Microsoft Windows, please consult the user manual for your operating system.

### **3.4. Starting RM5 Converter**

To start the Review Manager 5 data converter program (presently, however, for dichotomous outcomes only), double-click the *RM5Converter.jar* file.

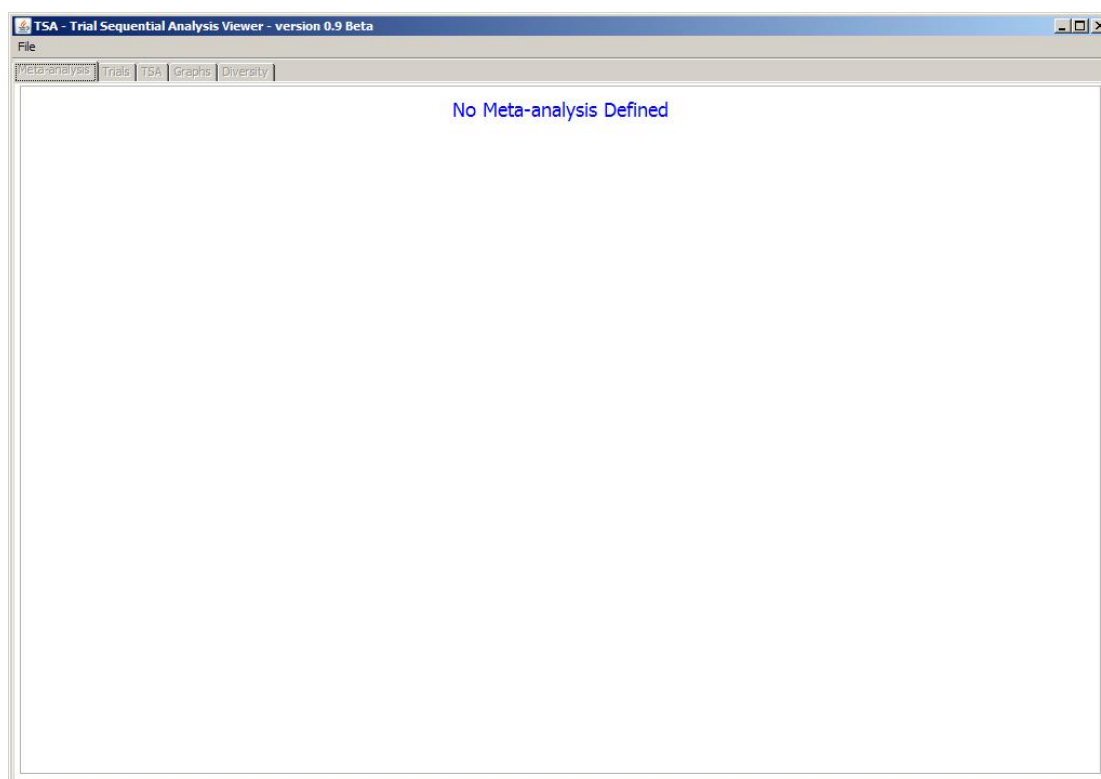
#### **3.4.1 Why doesn't RM5 start?**

The *RM5Converter.jar* has the same basic prerequisites as *TSA.jar*, so if you are having difficulty opening it, please consult section 3.3.1. Also, *RM5 Converter.jar* requires the *TSA.jar* file to be able to run. For this reason, *TSA.jar* has to be located in the same folder as *RM5Converter.jar*.

## 4. How to use TSA

### 4.1. Getting started

When TSA is started, a window similar to figure 13 should appear. The starting window should contain a menu bar with the menus *File*, *Settings*, and *Help*, as well as five greyed out (non-selectable) tabs: *Meta-analysis*, *Trials*, *TSA*, *Graphs*, and *Diversity*.



**Figure 13** The TSA starting window.

#### 4.1.1. Creating a new meta-analysis

To create a new meta-analysis, go to the menu bar and select *File > New Meta-analysis*. A dialogue box will appear (figure 14), allowing you to name your meta-analysis, choose the type of data that will be meta-analysed (dichotomous data or continuous data), define which two interventions are being compared, define whether the outcome type is 'negative' or 'positive' and add comments. Press *Create* to create the new meta-analysis. Press *Cancel* to cancel this action. If you want to edit the name of the meta-analysis, the interventions, or your comments, go to the menu bar and select *File > Edit meta-analysis*. The dialogue box shown in figure 14 should then re-appear.

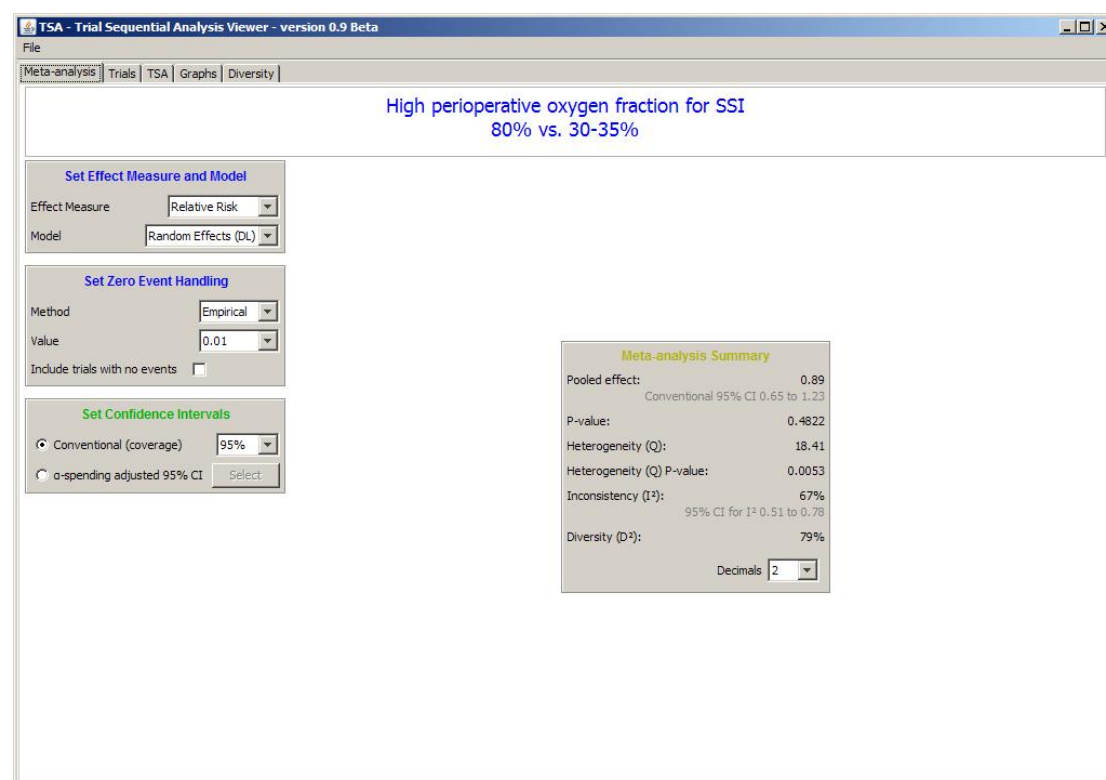
The image shows a software dialog box titled "New Meta-analysis". It contains the following fields and controls:

- Outcome section:**
  - Data Type:** A dropdown menu currently showing "Dichotomous".
  - Name:** An empty text input field.
- Comparison section:**
  - Label for Group 1:** An empty text input field.
  - Label for group 2:** An empty text input field.
  - Outcome type:** Two radio buttons, "Negative" (which is selected) and "Positive".
- Comments section:** A large, empty rectangular text area.
- Buttons:** "Create" and "Cancel" buttons at the bottom right.

**Figure 14** Dialogue box for creating a new meta-analysis.

*Note, binary data, negative outcomes are outcomes like mortality, stroke, or new cancer incidences; positive outcomes are outcomes like survival, clearance of a virus, or smoking cessation. For continuous data, negative outcomes are outcomes where an increase in the mean response is a bad thing (e.g., increase in depression score), and positive outcomes are outcomes where an increase in the mean response is a good thing (e.g., platelet count). The TSA software requires the designation of the outcome as negative or positive to determine which intervention arm the results favour.*

After creating your new meta-analysis, a number of options should appear in the left side of the starting window. (These options will be described in section 4.3. *Defining your meta-analysis settings.*)



**Figure 15** Starting window after a new meta-analysis has been created and data have been entered.

A box titled *Meta-analysis Summary* should appear in the middle of the window.

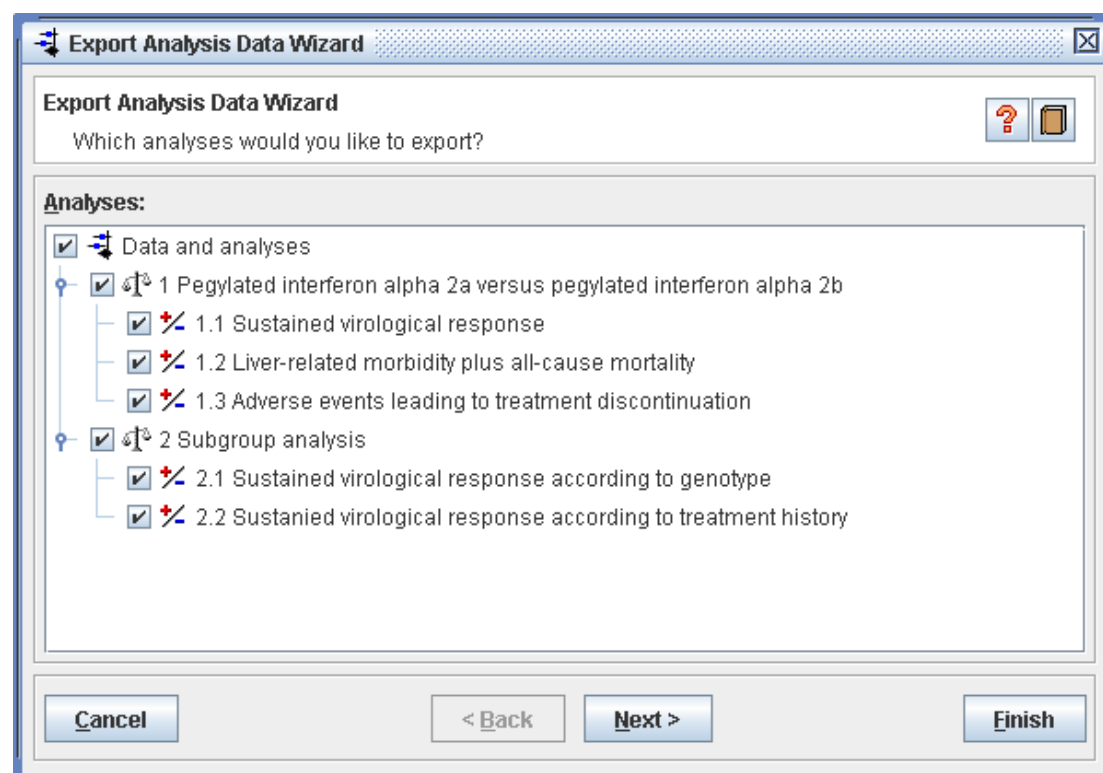
#### **4.1.2. Saving a TSA file and opening an existing TSA file**

If you wish to save your work, go to the menu bar and select *File > Save as...*

If you wish to continue working on an already created TSA file, go to the menu bar, select *File > Open*, and locate the TSA file on which you wish to continue working.

#### **4.1.3. Importing meta-analysis data from Review Manager v.5**

To import meta-analysis data saved in a Review Manager v.5 file (\*.rm5) (presently, however, for dichotomous outcomes only), you will need to use the separate software application *RM5Converter*, which is included in the Zip archive that you downloaded before installing TSA (see chapter 3).



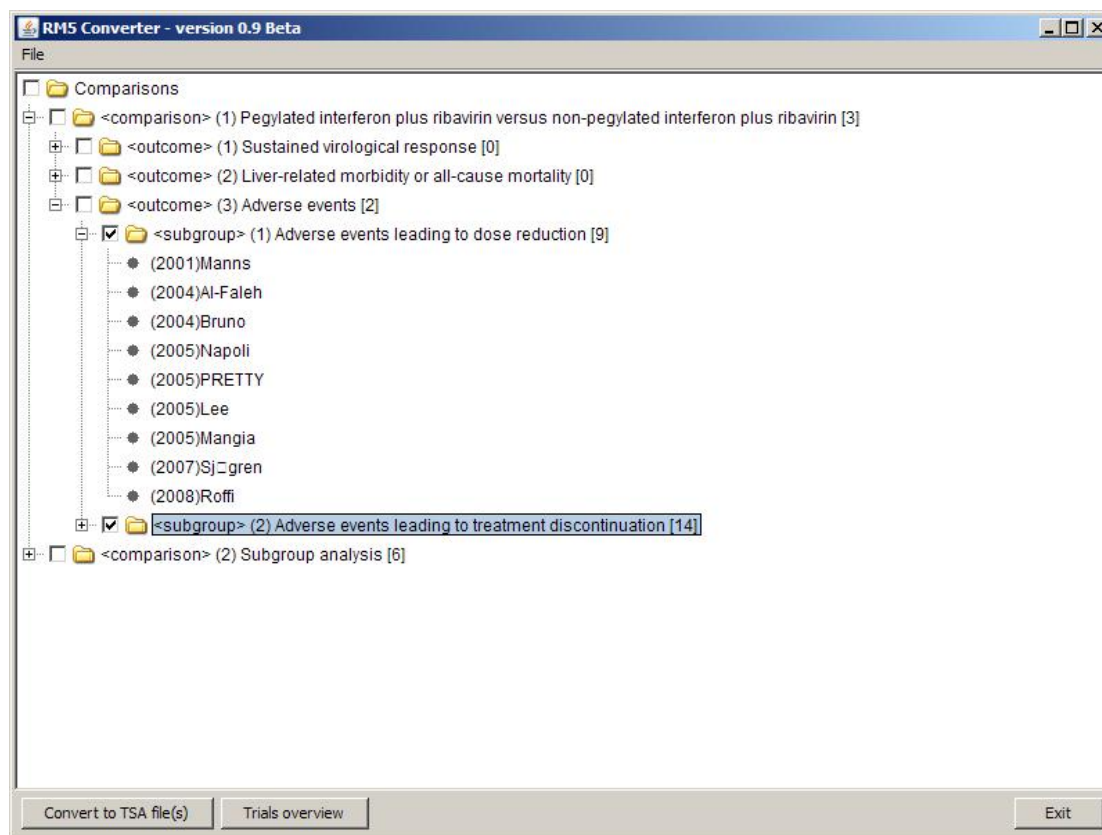
**Figure 16** Pop-up export analysis data wizard window in Review Manager v.5, which allows you to select the meta-analyses you wish to export (presently, however, only for dichotomous outcomes) as a \*.csv file. The RevMan file to be converted is from the Cochrane review 'Pegylated interferon alfa-2a versus pegylated interferon alfa-2b for treating chronic hepatitis C'.<sup>63</sup>

*RM5Converter* can read *comma separated files* (\*.csv). The first thing you need to do, therefore, is to convert your RevMan file into a comma separated file. Open your RevMan file in Review Manager v.5. in the menu-bar, select *File > Export > Data and analyses*. A pop-up window with a check box tree structure will appear (figure 16). Check the meta-analysis data that you wish to export as a comma separated file, and click on the *Next* button.

On the following screen check the three first checkboxes: Comparison Number, Outcome Number, and Subgroup Number. Then press *Finish*.

Note, if you click the *Next button* twice you will be presented with the option of choosing a field delimiter (what separates the cells in the data). It is important that the field delimiter is a comma (this is the default).





**Figure 17** Check the box tree structure in RM5Converter.

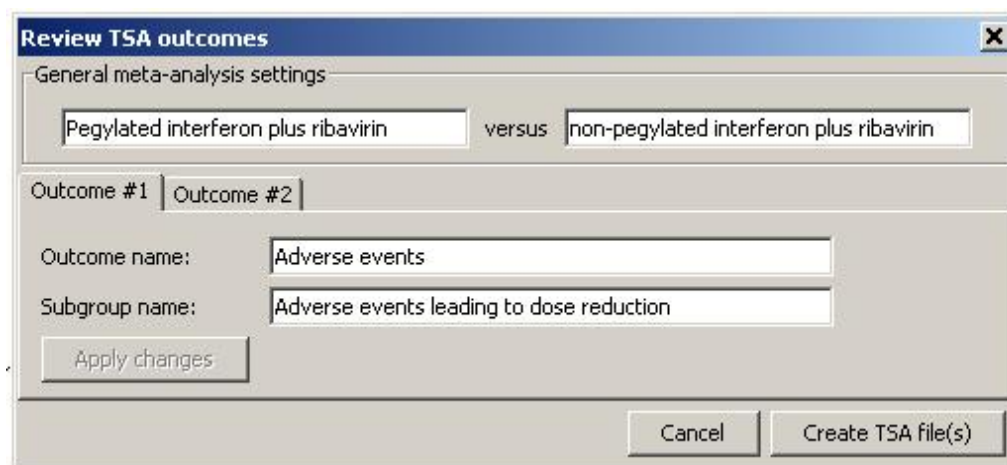
After you have exported your data to a \*.csv file, open *RM5Converter* by double-clicking on the icon. Go to the menu bar and select *File > Open*. A check box tree structure will appear in the application window (figure 17). The data is structured the same way as in Review Manager v.5. For each comparison, there can be multiple outcomes and each outcome represents a meta-analysis. If a meta-analysis contains subgroup analyses, the subgroups will be listed under each outcome. If a comparison is checked, all outcomes under that comparison will automatically be checked. Also, if an outcome is checked, all subgroups under that outcome will automatically be checked. All trials under each comparison, outcome, or subgroup are automatically included. You don't have to convert everything listed under a given comparison. You can 'uncheck' the comparisons, outcomes, and/or subgroups that you do not want to convert. If the trials in the subgroups are unique, you will be presented with the option of combining these subgroups into a single analysis.

The screenshot shows a window titled 'Trials Overview' with a 'Trial List' section. The list contains 25 trials, each with an identifier, a 'Low Bias Risk' checkbox (all are unchecked), and a 'Data' column showing 'Intervention: X/Y - Control: Z/W'.

Identifier	Low Bias Risk	Data
Derbala (2005)	<input type="checkbox"/>	Intervention: 5/35 - Control: 4/35
Izumi (2004)	<input type="checkbox"/>	Intervention: 10/23 - Control: 8/26
Roffi (2008)	<input type="checkbox"/>	Intervention: 17/24 - Control: 5/8
Cariti (2002)	<input type="checkbox"/>	Intervention: 41/60 - Control: 17/57
Nevens (2005)	<input type="checkbox"/>	Intervention: 120/230 - Control: 58/213
Lee (2005)	<input type="checkbox"/>	Intervention: 51/76 - Control: 49/77
Sjögren (2007)	<input type="checkbox"/>	Intervention: 12/29 - Control: 11/30
Bruno (2004)	<input type="checkbox"/>	Intervention: 31/163 - Control: 46/160
Horsmans (2008)	<input type="checkbox"/>	Intervention: 10/23 - Control: 8/26
Shobokshi (2003)	<input type="checkbox"/>	Intervention: 30/60 - Control: 18/60
PRETTY (2005)	<input type="checkbox"/>	Intervention: 19/89 - Control: 13/89
Manns (2001)	<input type="checkbox"/>	Intervention: 518/1025 - Control: 235/505
Esmat (2003)	<input type="checkbox"/>	Intervention: 1/100 - Control: 1/100
Rahman (1) (2007)	<input type="checkbox"/>	Intervention: 29/72 - Control: 21/72
Napoli (2005)	<input type="checkbox"/>	Intervention: 12/21 - Control: 3/19
Al-Faleh (2004)	<input type="checkbox"/>	Intervention: 6/48 - Control: 4/48
Dollinger (2005)	<input type="checkbox"/>	Intervention: 5/22 - Control: 4/18
Rahman (2,3) (2007)	<input type="checkbox"/>	Intervention: 88/131 - Control: 85/131
Mangia (2005)	<input type="checkbox"/>	Intervention: 14/121 - Control: 14/120
Fargion (2004)	<input type="checkbox"/>	Intervention: 22/92 - Control: 18/93
Wakil (2006)	<input type="checkbox"/>	Intervention: 12/18 - Control: 10/32
Thakeb (2003)	<input type="checkbox"/>	Intervention: 35/51 - Control: 8/49
Hinrichsen (2002)	<input type="checkbox"/>	Intervention: 24/36 - Control: 22/36
Derbala (2006)	<input type="checkbox"/>	Intervention: 25/40 - Control: 9/40
Scotto (2005)	<input type="checkbox"/>	Intervention: 3/26 - Control: 10/52

**Figure 18** Trials overview with bias risk check boxes.

If you click on the *Trials overview* button, a new window with a list of all the trials in the csv file will open. Each trial has an associated checkbox, indicating whether the trial is designated as a 'low bias-risk' trial or not (default is high bias-risk). You can change the designated bias risk of a trial by checking (or un-checking) its bias checkbox. Click on the *Close* button once you are done defining bias-risks.

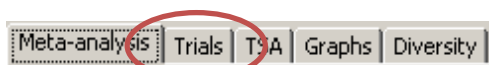


**Figure 19** Reviewing TSA outcomes in the RM5Converter.

Once you have checked all the comparisons, outcomes, and subgroups you want to include for your trial sequential analysis, click on the button in the bottom of the window titled *Create TSA file(s)*. A pop-up window, allowing you to review the names of your interventions, outcomes, and subgroups will appear. Once you have reviewed your selections, click on the button *Create TSA file(s)* and save your selected meta-analyses as TSA files in the specified folder.

## 4.2. Adding, editing, and deleting trials

Right below the menu bar in the TSA program, you will find five tabs: *Meta-analysis*, *Trials*, *TSA*, *Graphs*, and *Diversity*. To add, edit, or delete any trials in your meta-analysis, first select the *Trials* tab (figure 20).



**Figure 20** Click on the Trials tab when you want to add, edit, or delete trials in your meta-analysis.

In the left side of the window (in the Trials tab) there should be three areas: *Add Dichotomous/Continuous Trial*, *Edit/Delete Trial*, and *Ignore Trials*.

### 4.2.1. Adding trials

To add a new trial, fill in the input fields in the *Add Dichotomous/Continuous Trial* area. Regardless the type of data you are meta-analysing, you are required to provide some name or title for the study in the '*Study:*' input field

(typically the study acronym or the last name of the first author). You also need to provide the year that the study was published in the 'Year:' input field. You have the option to check the trial as a low bias risk trial.

If you are working with dichotomous data, you are required to enter the number of events and total number of patients in the (experimental) intervention group and the control group (figure 21).

Add Dichotomous Trial		Add Continuous Trial		
Study :	<input type="text"/>	Study :	<input type="text"/>	
Year :	<input type="text"/>	Year :	<input type="text"/>	
	Event      Total		Mean      Standard      Group	
			Response      Deviation      Size	
Intervention	<input type="text"/> <input type="text"/>	Intervention	<input type="text"/>	<input type="text"/> <input type="text"/>
Control	<input type="text"/> <input type="text"/>	Control	<input type="text"/>	<input type="text"/> <input type="text"/>
Low Bias Risk	<input type="checkbox"/>	Low Bias Risk	<input type="checkbox"/>	
Comment :	<input type="text"/>	Comment :	<input type="text"/>	
	<input type="button" value="Add Trial"/>		<input type="button" value="Add Trial"/>	

**Figure 21** Areas where you input the required data when adding a new dichotomous data trial (left) or continuous data trial (right).

If you are working with continuous data, you are required to enter the mean, standard deviation, and group size (number of patients) for the (experimental) intervention group and the control group (figure 21). It is also possible (but not necessary) to add a comment about the entered data. To submit the entered data, click on the *Add Trial* button.

In the right side of the window, you should find four columns: Study, Bias risk, Ignore, and Data. If you have added trials, a list of these trials should appear as in figure 22. The names and publication years of the added trials should appear in the first column (from the left) in the format '(year) title'. The assigned bias risks of the respective trials should appear in the second column. The bias risk of a trial can either be 'Low' (green letters) or 'High' (red letters). The third column gives you the option of ignoring one or more added

trial(s) for when you are performing your meta-analyses. Simply check the 'ignore' check box to ignore a trial. The fourth column should display the trial data. For dichotomous trials the format is 'Intervention: Events/Total. Control: Events/Total'. For continuous data the format is 'Intervention: Mean Reponse/Standard Deviation/Group Size. Control: Mean Reponse/Standard Deviation/Group Size'.

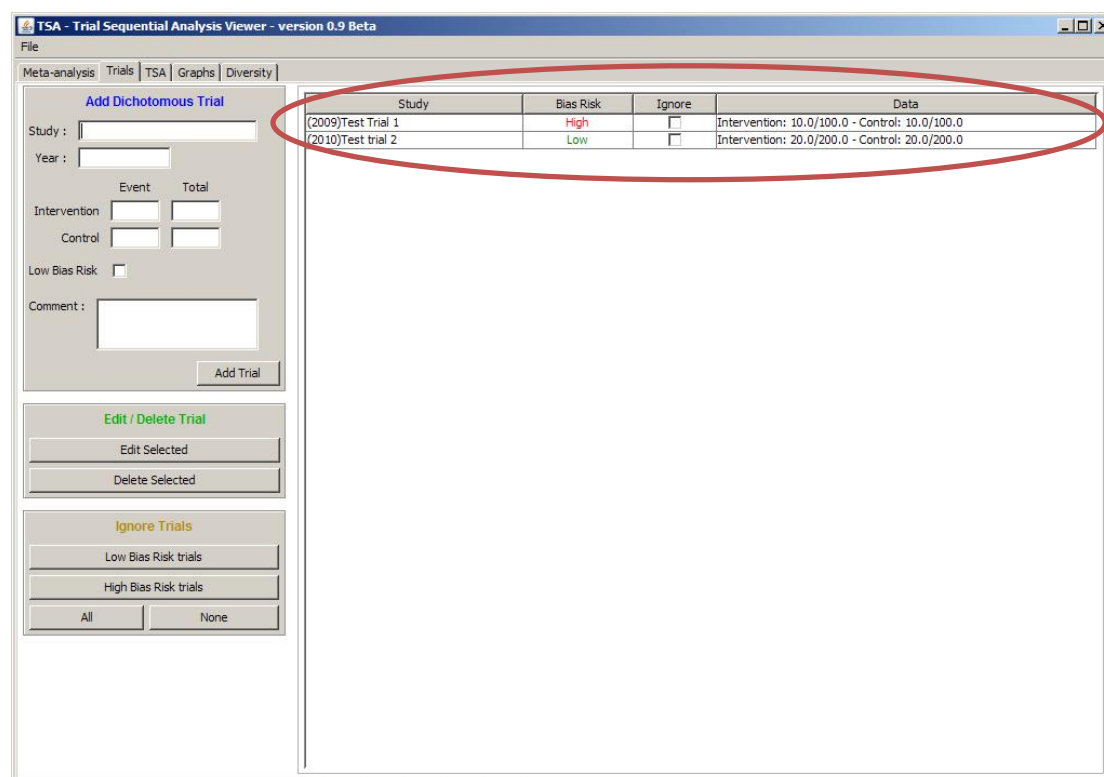


Figure 22 List of added trials marked within the red ellipse.

#### 4.2.2. Editing and deleting trials

To edit trial data, first select the row for the trial you wish to edit and then click on the Edit Trial button in the *Edit/Delete Selected Trial* area (figure 23). Alternatively, you can double click on the row for the trial you wish to edit.

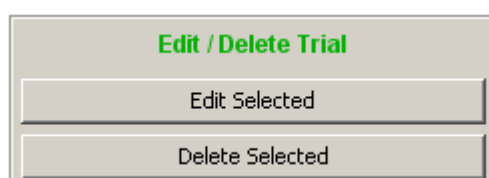


Figure 23 The Edit/Delete Selected Trial area.

The trial data appears in the same area where you type in new trial data and you can now edit data. This area will now contain an *Edit Trial* button instead of an *Add Trial* button. To edit the trial data, change the content in the fields you want to edit and click on the *Edit Trial* button.

If you wish to delete a trial, select the row for the trial you wish to delete, and press the *Delete Trial* button in the *Edit/Delete Selected Trial* area. Alternatively, you can select the row for the trial you wish to delete and press the <Delete> button on your keyboard.

### 4.3. Defining your meta-analysis settings

The TSA program provides a number of options for performing meta-analysis. You can choose between a number of effect measures, statistical models, zero-event data handling methods (for dichotomous data), and confidence interval coverage levels. All of these options can be set in the *Meta-analysis* tab to the left of the *Trials* tab (figure 24).

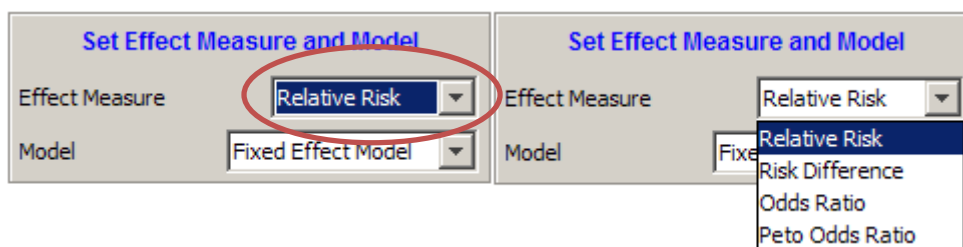


**Figure 24** Click on the Meta-analysis tab when you want to set your effect measure, statistical model, or zero-event handling method.

In the left side of the window you will find the *Set Effect Measure and Model* area, the *Set Zero Event Handling* area, and the *Set Confidence Intervals* area (figures 25-28). In the middle of the window, you will find the *Meta-analysis Summary* area.

#### 4.3.1. Choosing your association measure

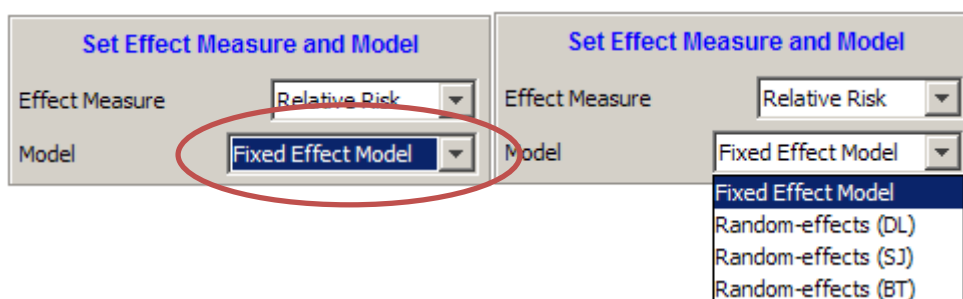
The TSA program provides the same effect measures as *Review Manager version 5* (see section 2.1.1. for a description of these measures). To select an effect measure, first click on the Effect Measure drop-box in the *Set Effect Measure and Model* area in order to display the available effect measures (figure 25, marked area in the left side picture), then click on the effect measure you wish to use for your meta-analysis.



**Figure 25** Select effect measure by clicking on the *Effect Measure* drop-box.

### 4.3.2. Choosing your statistical model

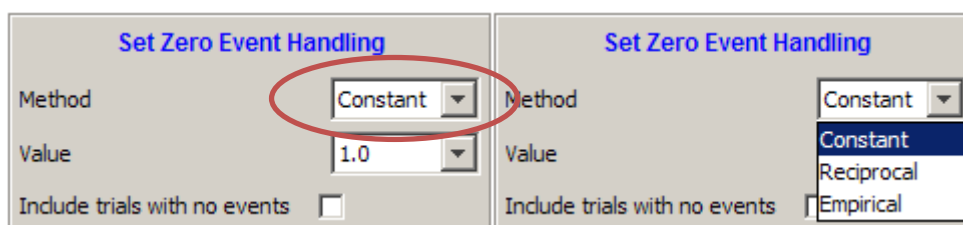
The TSA program provides four statistical models for pooling meta-analysis data – three of which are variants of the random-effects model (see section 2.1.2). To set your statistical model, first click on the *Model* drop-box to display the available effect measures (figure 26, marked area in the left side picture), and then click on the model you wish to use for your meta-analysis.



**Figure 26** Select effect measure by clicking on the *Model* drop-box.

### 4.3.3. Choosing a method for handling zero-event data

The TSA program provides three methods for handling zero-event data (see section 2.1.4). To select the method you wish to employ for handling zero-event data, first click on the *Method* drop-box to display the available continuity correction methods, and then click on the method you wish to employ (figure 27).

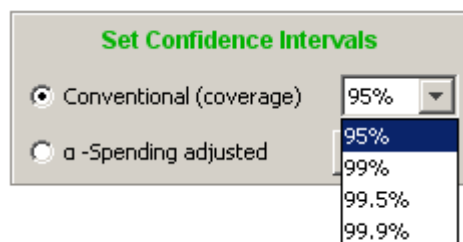


**Figure 27** Select continuity correction method by clicking on the *Method* drop-box.

You also need to set the continuity correction factor. In the TSA program, the correction factors are derived from sum of the correction factors in the two groups (also referred to as the 'Value'). For example, the sum of the correction factors in the continuity correction used in Review Manager is  $1=0.5+0.5$  - because 0.5 is added to the number of events in both groups. To set the sum of two correction factors, first click on the *Value* drop-box, then select the sum you wish the two correction factors to add up to. In addition, you have the option of applying continuity correction on trials that have zero events (or non-events) in both arms. To do so, check the box titled 'Include trials with no events'

#### 4.3.4. Choosing the type of confidence interval

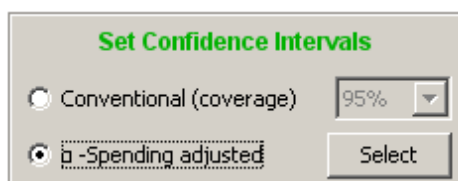
TSA provides a number of options for the type of confidence interval you wish to employ (figure 28). If you are employing conventional confidence intervals you can choose between coverage levels 95%, 99%, 99.5%, and 99.9%. To do so, check the 'Conventional (coverage)' radio button left in the *Set Confidence Intervals* area, click on the drop down box to the right and select your desired coverage.



**Figure 28** Choose you coverage for conventional confidence intervals.

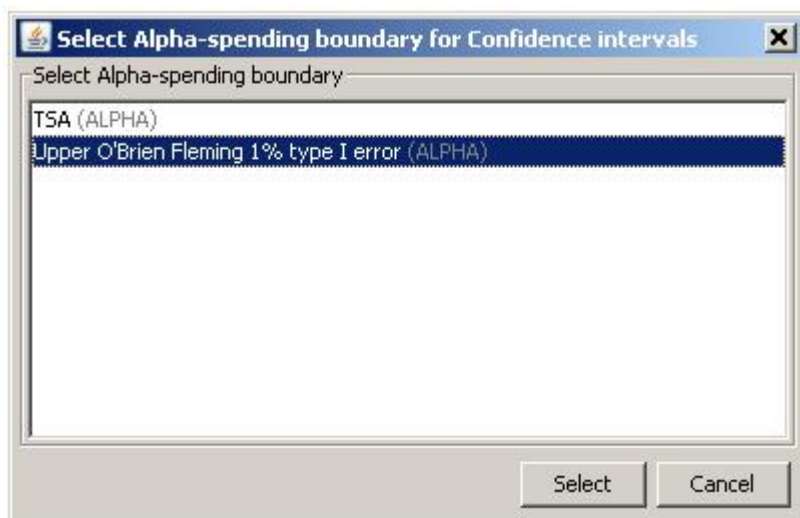
If you have already constructed adjusted significance test boundaries using an  $\alpha$ -spending function (see section 2.2.4 and 4.4.1), you will also have the option of obtaining the  $\alpha$ -spending adjusted confidence interval (see section 2.2.5). To do so, first click on the 'α-spending adjusted' radio button in the *Set Confidence Intervals* area and subsequently click on the 'Select button' (figure 29).





**Figure 29** Select  $\alpha$ -spending function adjusted confidence intervals.

A pop-up window with a list of your added alpha-spending boundaries should appear in the middle of the screen. Select which of the alpha-spending boundaries the adjustment should be based on and click on the Select button (figure 30). Note, the cumulative coverage of the alpha-spending adjusted confidence intervals will correspond to the alpha level set for the chosen alpha-spending function.



**Figure 30** Choose the alpha-spending boundaries on which the adjustment should be based.

Also note that only  $\alpha$ -spending boundaries that can be calculated and have not been 'ignored' will be included in the list (see section 4.4.1).

#### **4.4. Applying adjusted significance tests (applying TSA)**

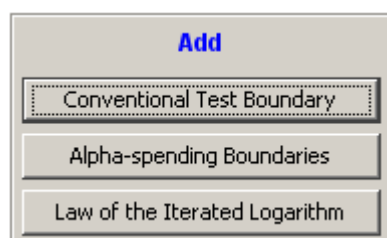
TSA currently provides two methods for adjusted significance testing. These are the *O'Brien-Fleming  $\alpha$ -spending method*, described in section 2.2.4., and the *law of the iterated logarithm method*, described in section 2.2.6. TSA also provides the option to combine the *O'Brien-Fleming method* with futility testing as described in section 2.2.7. To apply these methods, click on the *TSA* tab (to the right of the *Trials* tab) as shown in figure 31.



**Figure 31** Click on the TSA tab when you want apply methods for adjusted significance testing.

#### **4.4.1. Adding a significance test**

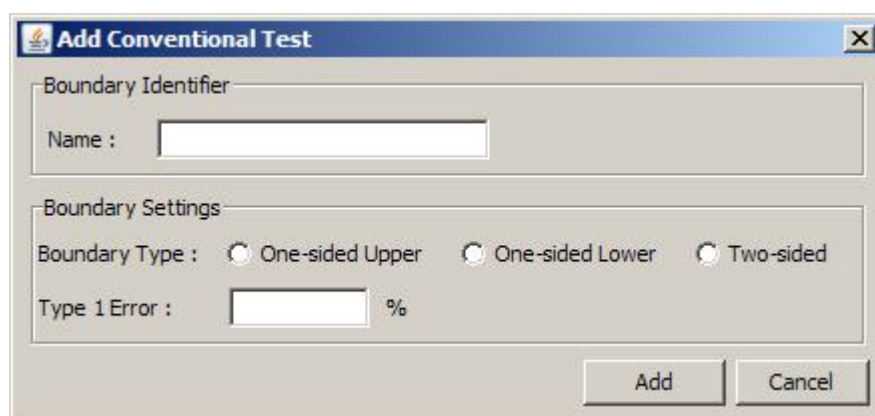
In the upper left side of the window, you will find the *Add* area (figure 32), which contains the buttons *Conventional Test Boundary*, *Alpha-spending Boundaries*, and *Law of the Iterated Logarithm*. When you click on any of these three buttons a new window should appear in the middle of the TSA program window. This window should contain a number of fields, which will allow you to define the settings for the type of significance test you wish to apply.



**Figure 32** Click on one of the buttons to add a new significance test.

##### *The conventional significance boundary*

The *Conventional* option allows you to add a boundary for the Z-curve which corresponds to a single significance test with some maximum type I error risk,  $\alpha$ . For example, a *conventional* boundary for a two-sided  $\alpha=5\%$  single significance test will produce two horizontal lines at 1.96 and -1.96. When you click on the *Conventional* button, a window similar to the one shown in figure 33 should appear.



**Figure 33** *Conventional Test* setting pop-up window that appears when clicking on the *Conventional Test Boundary* button.

You will need to give your conventional test a name (e.g., 'single test 5% threshold'), define whether your test is two-sided (symmetric) or one-sided and what your overall (single test) maximum type I error will be. For one-sided tests, the *Upper* one-sided test will only test for superiority of the experimental intervention, whereas the *Lower* will only test for superiority of the control intervention. For binary data meta-analysis, it should be noted that when the outcome is defined as a 'positive' rather than 'negative', (see section 4.1.1) the functions of *Upper* and *Lower* are reversed. When you have named your conventional boundary and defined the settings, press the *Add* button to add the boundary.

#### *The $\alpha$ -spending boundaries*

The *alpha-spending* option allows you to add adjusted significance boundaries for the Z curve with the  $\alpha$ -spending method described in section 2.2.4. Because the  $\alpha$ -spending method cannot be applied without determining some required meta-analysis information size, the information size calculations must be defined simultaneously. Therefore, the  $\alpha$ -spending boundaries setting window for dichotomous data meta-analysis will be different from continuous data meta-analysis with respect to the settings for the information size calculation. For dichotomous data meta-analysis, the  $\alpha$ -spending boundaries setting window that appear when you click on the *alpha-spending* button should similar to the one shown in figure 34.

**Add Dichotomous Alpha-spending Boundary**

Boundary Identifier

Name:

Hypothesis Testing

Boundary Type:  One-sided Upper  One-sided Lower  Two-sided

Type 1 Error:  %

$\alpha$ -spending Function:

Information Axis:  Sample Size  Event Size  Statistical Information

Inner Wedge

Apply Inner Wedge:

Power:  %

$\beta$ -spending Function:

Required Information Size

Information Size:   User Defined  Estimate

Type 1 Error:  %

Power:  %

Relative Risk Reduction:  %  User Defined  Low Bias Based

Incidence in Intervention arm:  %  User Defined

Incidence in Control arm:  %

Heterogeneity Correction:  %  User Defined  Model Variance Based

Add Cancel

**Figure 34** *Alpha-spending* boundaries setting pop-up window for dichotomous data meta-analysis that appears when clicking on the *alpha-spending* button.

For continuous data meta-analysis, the *alpha-spending* boundaries setting window that appear when you click on the *alpha-spending* button should be similar to the one shown in figure 35.

**Add Continuous Alpha-spending Boundary**

Boundary Identifier

Name:

Hypothesis Testing

Boundary Type:  One-sided Upper  One-sided Lower  Two-sided

Type 1 Error:  %

$\alpha$ -spending Function:

Information Axis:  Sample Size  Event Size  Statistical Information

Inner Wedge

Apply Inner Wedge:

Power:  %

$\beta$ -spending Function:

Required Information Size

Information Size:   User Defined  Estimate

Type 1 Error:  %

Power:  %

Mean Difference:   User Defined  Empirical  Low Bias

Variance:   User Defined  Empirical  Low Bias

Heterogeneity Correction:  %  User Defined  Model Variance Based

Add Cancel

**Figure 35** Alpha-spending boundary setting pop-up window for continuous data meta-analysis that appears when clicking on the *alpha-spending* button.

First, you will need to give your  $\alpha$ -spending based test a name (e.g., '5% symmetric O'Brien-Fleming'). You will then need to define if you wish to employ a two-sided (symmetric) or one-sided test, what your overall maximum type I error will be, what type of  $\alpha$ -spending you wish to employ (currently only the O'Brien-Fleming function is available). You will then need to decide whether you wish to define the information in your meta-analysis as the accumulated number of patients (sample size), accumulated number of events (event size), or accumulated statistical information. Again, for one-sided tests the *Upper* one-sided test will only test for superiority of the

experimental intervention, whereas the *Lower* will only test for superiority of the control intervention. For binary data meta-analysis, it should be noted that when the outcome is defined as a 'positive' rather than a 'negative' outcome (see section 4.1.1) the functions of *Upper* and *Lower* are reversed.

To test for futility (i.e., apply *inner wedge* futility boundaries) check the 'Apply Inner wedge' checkbox. The type II error (or power) for the futility boundaries will automatically be set when you enter your settings for your information size calculation (see below). Currently, the only  $\beta$ -spending function available in TSA is the O'Brien-Fleming function.

You will need to input the necessary components for the required information size calculation. You will have the option to define the required information as any arbitrary number you may have obtained independent of the TSA software. To submit your own value for IS, check the radio button 'User defined' and type in the required IS. You also have the option to estimate the required IS according to the methods delineated in section 2.2.1. To use TSA to calculate the required IS, check the radio button 'Estimate'. The required IS estimate will automatically be generated with respect to the type of information you are accumulating. For example, if you selected 'sample size' under 'Information Axis', the required information size will be generated as the required number of patients in the meta-analysis.

The IS calculation will automatically be based on the maximum type I error you defined for the  $\alpha$ -spending boundary, but you will need to enter your desired maximum type II error/minimum desired power (1-type II error) into the input field 'Power'.

You have two options for adjusting the required information size for heterogeneity in the meta-analysis. The first option is to base the heterogeneity adjustment on the estimated ratio between the variance in the selected random-effects model and the variance in the fixed effect model (see section 2.2.1). To use this option, check the radio button 'Model Variance

Based'. Note that if you have selected the fixed-effect model, this adjustments factor is always equal to 1, and thus, no adjustment is applied.

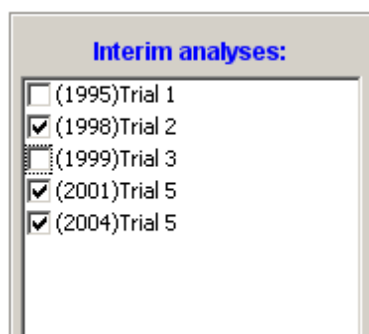
The second option is to make a guess estimate of predicted heterogeneity. When your meta-analysis includes an insufficient number of trials to reliably estimate the adjustment factor, you may adjust the required information size for some *a priori* maximum or plausible anticipated degree of heterogeneity. To use this option, check the 'User Defined' radio button and type in the maximum anticipated heterogeneity in the input field to the left. Here heterogeneity is defined as the percentage of the total variance in the meta-analysis which is explained by between-trial variation rather than within-trial variation. Thus, a user defined adjustment of 50%, for example, yields a required information size that allow for reliable inference when approximately half of the total variation among trial in the meta-analysis is explained by the between-trial variation.

To set the anticipated event rates and intervention effect for a dichotomous data meta-analysis, you only need to fill in two of the three fields: 'Relative risk reduction', 'Incidence in Intervention Group', and 'Incidence in Control Group'. If you have categorized some of your included trials as low-bias risk trials, you may use the pooled meta-analysis estimates of these trials as your anticipated relative risk reduction. To use this option, select the 'Low-bias Based' option.

To set the anticipated mean difference and variance for a continuous data meta-analysis, you only need to fill in two fields: 'Mean difference' and 'Variance'. If you have categorised some of your included trials as low-bias risk trials, you may use the pooled meta-analysis estimates of these trials as your anticipated mean difference and variance, again by selecting 'Low-bias Based' option. You also have the option to use the pooled estimate of all included trials (regardless of bias risk) as your anticipated variance. To use all trials, select the 'Empirical' option.

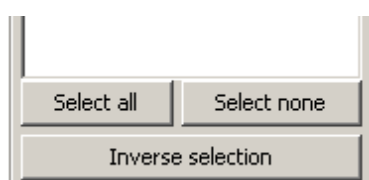
When you have named your  $\alpha$ -spending boundaries, defined the hypothesis test settings, and the parameters for your information size calculation, press the *Add* button to add the boundaries.

After you have added your  $\alpha$ -spending boundaries you will need to define when the meta-analysis was previously subjected to significance testing. Go to the *Interim looks* to the right of the list of adjusted significance tests and check (or uncheck) the trials after which significance testing were previously performed. In figure 36, trials 2, 4, and 5 have been checked, and trials 1 and 3 have been unchecked, meaning that three meta-analyses (including significance testing) were performed over time: one including trial 1 and 2, one including trials 1 to 4, and one including trials 1 to 5. Note that the last trial on the list should always be checked, as this represents the significance test you are employing on all included trials.



**Figure 36** *Alpha-spending* boundary setting pop-up window for continuous data meta-analysis that appears when clicking on the *alpha-spending* button.

In some cases, you may wish to check or uncheck all trials for previous significance tests. Click on the 'Select none' button in the bottom of the *Interim analyses* area to uncheck all trials, or click on the 'Select all' button to check all trials (figure 37). In addition, you have the option to inverse the selection of interim looks.



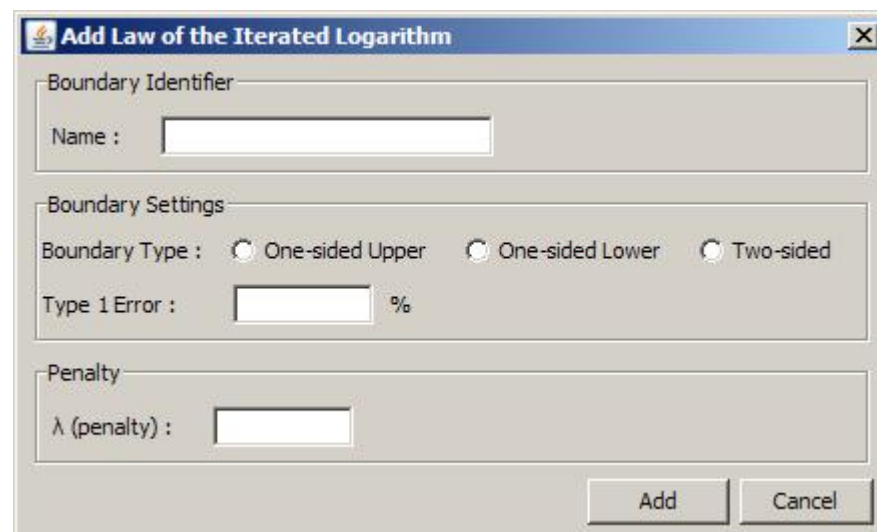
**Figure 37** Check or uncheck all trials for previous significance tests.



### *The law of the iterated logarithm penalised Z curve*

The *law of the iterated logarithm* option allows you to perform adjusted significance testing by penalising the Z curve with the methods described in section 2.2.6. When you click on the *Law of Iterated Logarithm* button, a window like the one shown in figure 38 should appear.

You will need to give your Z curve penalisation a name (e.g., '5% symmetric LIL'), define whether your test is two-sided (symmetric) or one-sided, what your overall maximum type I error will be, and set your penalisation parameter,  $\lambda$  (see section 2.2.6 and table 2). For one-sided tests, the *Upper* one-sided test will only test for superiority of the experimental intervention, whereas the *Lower* will only test for superiority of the control intervention. For binary data meta-analysis, it should be noted that when the outcome is defined as a 'positive' rather than a 'negative' outcome (see section 4.1.1) the functions of *Upper* and *Lower* are reversed.



**Figure 38** *Law of the iterated logarithm* penalisation setting pop-up window that appears when clicking on the *Law of Iterated Logarithm* button.

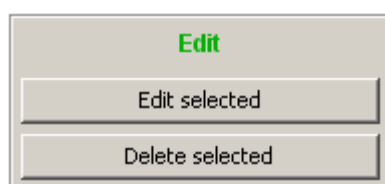
#### **4.4.2. Editing and deleting a significance test**

Whenever a significance test is added it should appear in the middle of the screen. Each significance test you add will be represented by a row as shown in figure 39.

Identifier	Ignore	Type
TSA (ALPHA)	<input type="checkbox"/>	Alpha-spending
Conventional 2-sided (CON)	<input type="checkbox"/>	Conventional
LIL 5% (LIL)	<input type="checkbox"/>	Law of Iterated Logarithm

**Figure 39** List of added significance tests.

To edit a significance test, first select the row for the test you wish to edit and then click on the 'Edit selected' button in the *Edit* area (figure 40). Alternatively you can simply double click on the row for the test you wish to edit.



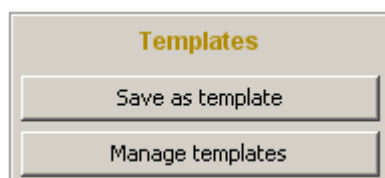
**Figure 40** The Edit/Delete Selected significance test.

The pop-up window with the test's settings will now appear. Make your edits and click on the 'Apply changes' button in the lower right corner of the pop-up window.

If you wish to delete a test, select the row for the test you wish to delete, and press the *Delete Selected* button in the *Edit* area. Alternatively you can select the row for the test you wish to delete and press the <Delete> button on your keyboard.

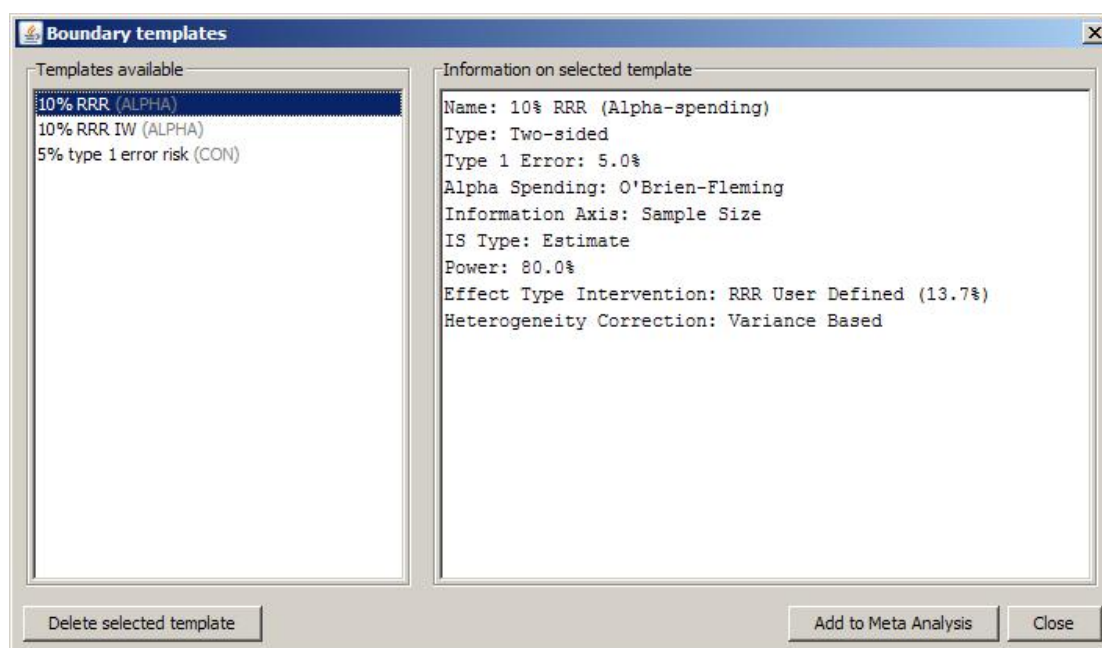
#### **4.4.3. Adding and loading significance test templates**

The *Templates* area in the lower left corner of the TSA tab provides you with the option of saving your constructed significance tests, and loading previously constructed significance tests (figure 41). If you wish to re-use a significance test for other meta-analyses, you can save this in your templates and load it at any other time. To save a constructed significance test as a template, select the row for the test you wish to save and click on the 'Save as template' button. To load a previously saved template, first click on the 'Manage templates' button.



**Figure 41** Template area where you can load and save (add) constructed significance tests.

A pop-up window will appear in the middle of TSA program window (figure 42). The list of available templates is shown to the left.



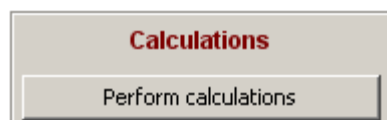
**Figure 42** Templates window. The significance test '10% RRR' has been selected and the settings of this test are displayed under 'Information on selected boundary'.

You can click on a template title to display the available significance tests' settings on the right side. To load a template significance test for your meta-analysis, select the template you wish to load and click on the 'Add to Meta-analysis' button. If you wish to delete one of the available templates permanently, select the template you wish to delete and click on the 'Delete selected Template' button.

#### **4.4.4. Performing the significance test calculations**

Once you have added all the significance tests you wish to employ, you need the TSA program to perform the necessary calculations. To achieve this, click

on the 'Perform calculations' button in the *Calculations* area under the *Edit* area. Depending on how many significance test you have added, the TSA program might take a few seconds to complete the calculations. O'Brien-Fleming ( $\alpha$ -spending) type boundaries with many interim analyses can take 5-10 seconds per set of boundaries to compute.



**Figure 43** Perform calculations button.

In some instances, there is such a small relative increase in information between two interim analyses that the numerical calculations (numerical integration of extremely small tail probabilities) for the  $\alpha$ -spending boundaries break down. For example, if the required information size is 20,000 patients and the interim analyses are performed after each trial, adding a new trial with 40 patients would only provide a 0.2% increment in the cumulative information fraction. To avoid breakdowns in the calculations, the TSA program automatically removes (un-checks) interim analyses that correspond to an information fraction increment of 1% or smaller. When this happens, a window will automatically pop up in the middle of the TSA program window to inform which interim analyses were removed (figure 44). The data of these trials are, however, retained in your TSA meta-analysis and in the cumulative Z-value.



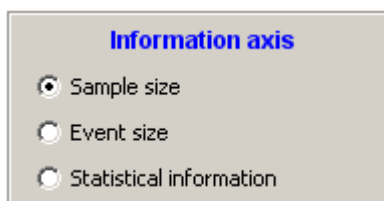
**Figure 44** Pop-up window that inform which interim analyses were removed. The data of these trials are, however, retained in your TSA meta-analysis and in the cumulative Z-value.

If you have added more than one significance test and do not wish to perform the calculations for all of these, you have the option to ignore significance tests. To ignore a significance test, check the checkbox in the mid column for the row corresponding to the significance test(s) you wish to ignore.

Identifier	Ignore	Type
TSA (ALPHA)	<input type="checkbox"/>	Alpha-spending
Conventional 2-sided (CON)	<input checked="" type="checkbox"/>	Alpha-spending boundary ignored by user
LIL 5% (LIL)	<input type="checkbox"/>	Law of Iterated Logarithm

**Figure 45** Example of an ignored significance test ('Conventional 2-sided' ignored).

The cumulative Z-curve and the significance boundaries (for  $\alpha$ -spending functions) can be displayed using one of three variables on the x-axis: sample size, the event size, or the statistical information. Significance tests defined on different scales cannot be displayed simultaneously in a graph, so you need to select one of these variables for the whole analysis. Check the appropriate radio button in the *Information Axis* area below the *Calculation* area (figure 46).



The image shows a rectangular box with a light grey background. At the top, the text "Information axis" is written in blue. Below this, there are three radio button options, each with a small circle to its left. The first option is "Sample size" and its radio button is selected (filled with a black dot). The second option is "Event size" and its radio button is unselected (empty). The third option is "Statistical information" and its radio button is unselected (empty).

**Figure 46** Radio buttons for choosing the information scale on which the cumulative significance testing is displayed.

If one or two of the three scales (sample size, event size, or statistical information) have not been selected in any of the added  $\alpha$ -spending boundaries, they will automatically be greyed out in the *Information Axis* area.

#### 4.5. Graphical options for TSA

The *Graph* option in the TSA program allows you to display the Z-curve and your constructed significance tests in relation to the strength of evidence (i.e., accumulated number of patients, events, or statistical information). It also provides a number of graphical editing options that may be useful when you

are preparing graphs for your article manuscripts. To go to the *Graph* option, click on the *Graph* tab (to the right of the *TSA* tab) as shown in figure 47.



**Figure 47** Click on the *Graphs* tab to view the Z-curve and your constructed significance tests displayed in relation to the strength of evidence.

In the left side of the TSA program window you will find the *Tests and boundaries Layout* area, the *Set Graph Layout*, and two print options ('Print current graph', and 'Generate TSA Report'). To the right of these areas you will find the graph displaying your Z-curve and constructed significance tests.

In the *Tests and boundaries layout* area you will find a number of graphical editing options that allow you to change the presentation of the Z-curve and your constructed significance tests (boundaries). Your constructed significance tests and the Z-curve will be listed in the white area; see figure 49. To change the presentation of one of these, first select one of the tests (curves) on the list and edit according to your preferences.

In the TSA program, you will have the option to edit the colour, the line type, and the type and size of the icon displayed at each trial or interim analysis, as well as the size and font of the test associated with a curve or a test. You also have the option to hide a curve/test from the graph.

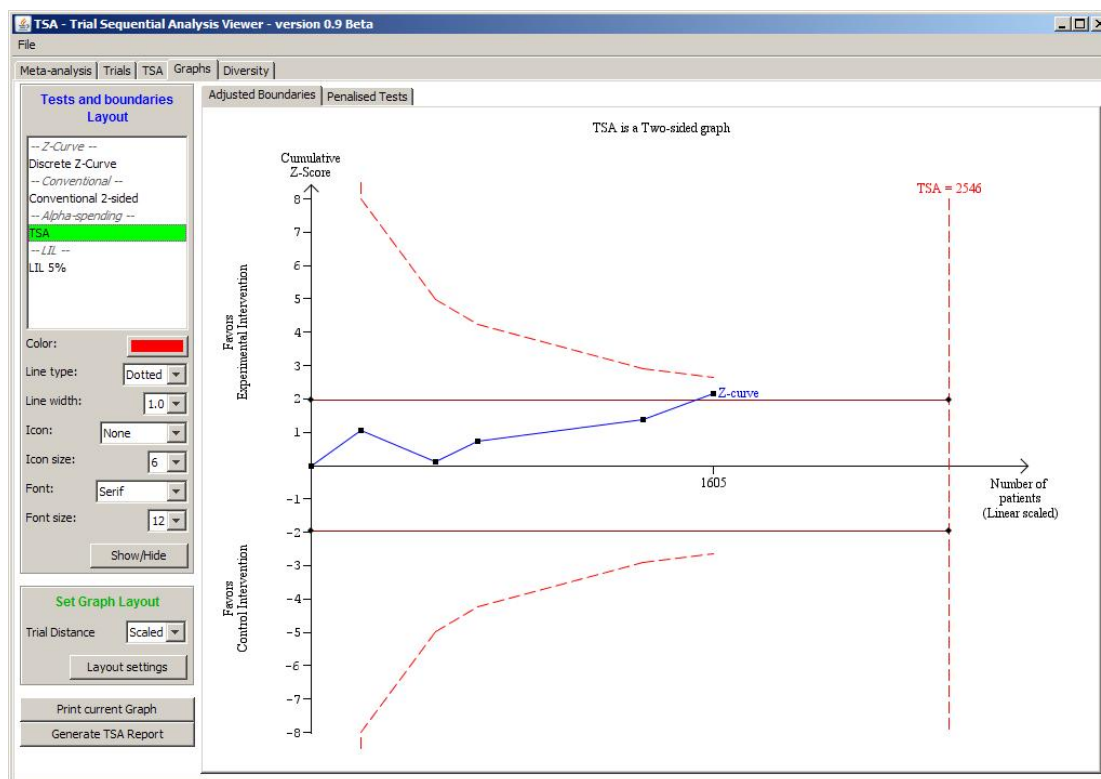


Figure 48 The Graph window.

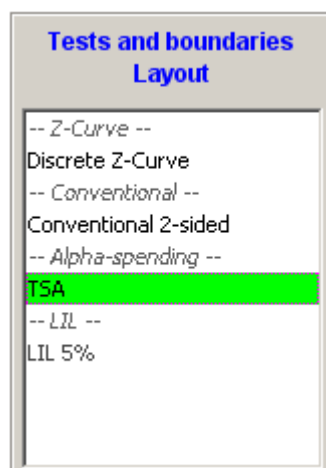
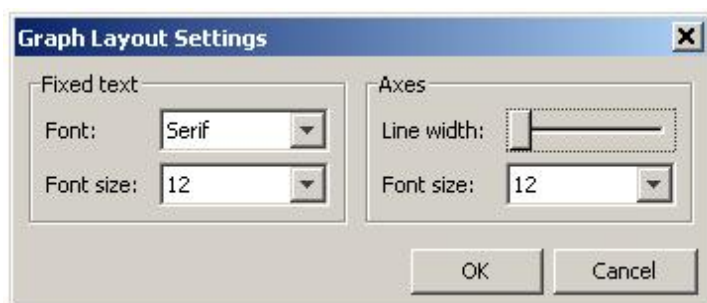


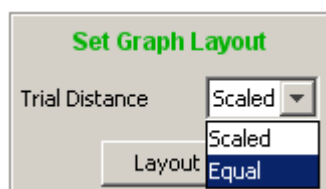
Figure 49 The significance tests and Z-curve listings area

In the *Set Graph Layout* area, you will find a number of options for changing the general graph presentation. If you click on the 'Layout settings' button, a pop-up window will appear (figure 50), providing you with the options of adjusting the width of the x-axis and y-axis, the coordinate font size, or the font and the size of the fixed text components on the graph.



**Figure 50** General graph layout settings to adjust fixed text components' font and font size, the width of the x-axis and y-axis, and the coordinate font size.

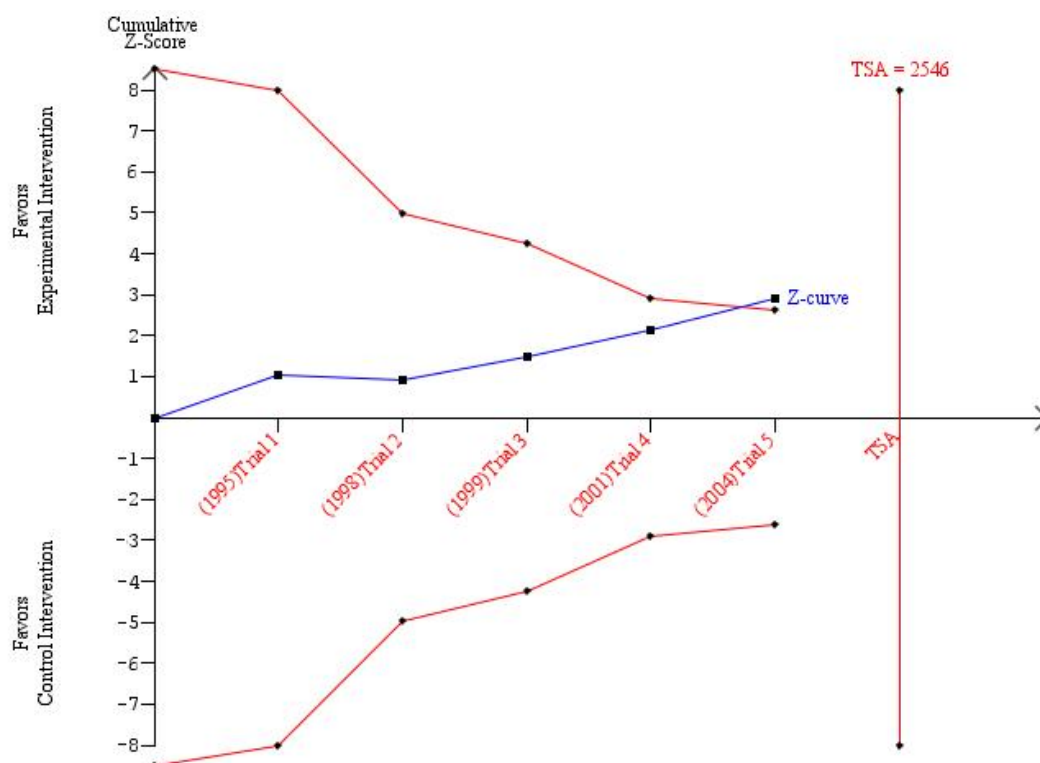
On the information axis, the distance between boundaries and the distance between the Z-values are conventionally displayed with respect to the relative increase in information. The TSA program automatically displays these distances in this scaled manner.



**Figure 51** Select 'Equal' distance for equal distance between trials on the x-axis.

In some instances, however, other layouts may provide a better basis for visual interpretation. The TSA program also provides the layout format used in the paper by Pogue and Yusuf, which displays trials at equal distance on the information axis and displays the trial titles at a 45° angle below the x-axis. To choose this layout format, click on the 'Trial Distance' drop down box in the *Set Graph Layout* area and select 'Equal' (figure 51).





**Figure 52** Select information-axis scaling/display format.

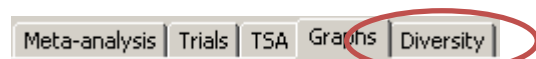
Adjusted significance tests based on  $\alpha$ -spending functions are, in effect, adjusted thresholds for the Z-curve, whereas adjusted significance tests based on the law of the iterated logarithm penalties are, in effect, adjusted test statistics that should be interpreted in relation to single-test significance test thresholds. Thus, combining these two approaches in one graph is not meaningful. The TSA program provides separate graphs for adjusted significance tests based on  $\alpha$ -spending functions and the law of the iterated logarithm penalties. To see the graphical representation of the calculated  $\alpha$ -spending boundaries, select the *Adjusted Boundaries* tab above the graph. To see the graphical representation of the calculated law of the iterated logarithm penalties, select the *Penalised Tests* tab above the graph (figure 53).



**Figure 53** View boundaries test or penalised test graph.

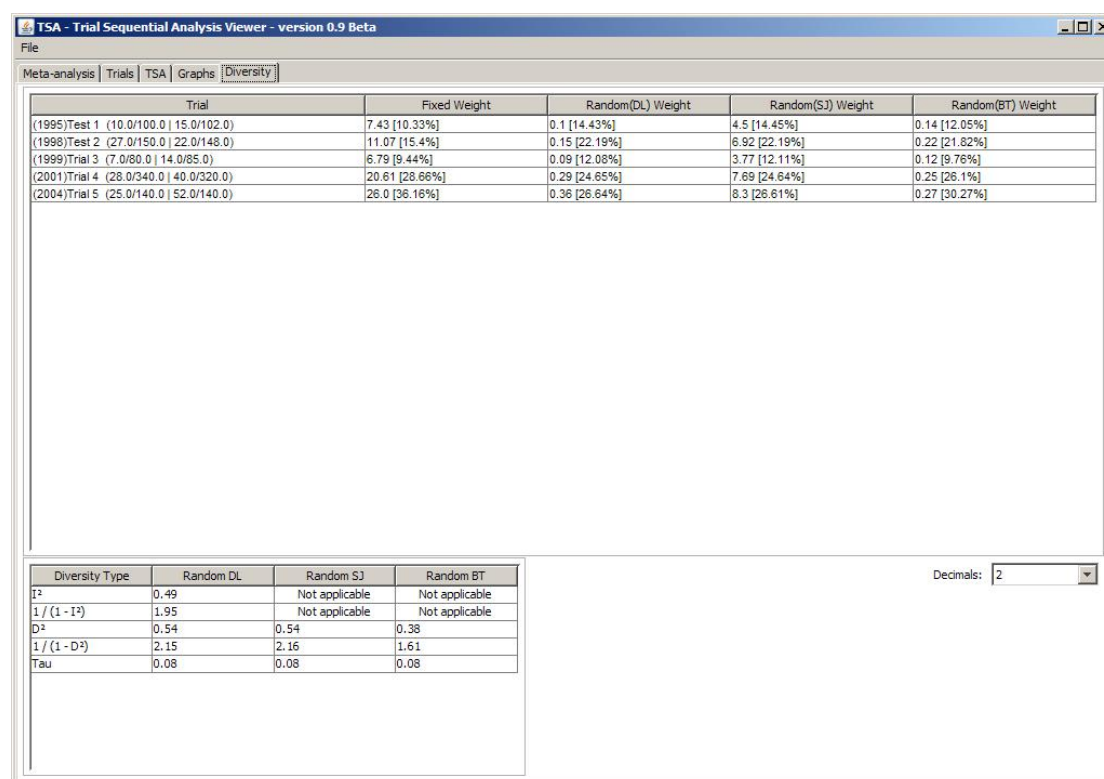
#### 4.6. Exploring diversity across trials

The TSA program also provides an option for exploring diversity estimates and comparing weights across the three random-effects models: DL, SJ, and BT. These options are available in the *Diversity* tab (figure 54)



**Figure 54** Click on the Diversity tab to explore diversity estimates and compare weights across random-effects models.

After you click on the diversity tab, a screen similar to the one shown in figure 55 should appear. In the upper part of the screen, the weights and weight percentages for each trial (rows), using each of the available models (columns), are displayed in the lower left corner. The following things are displayed for each of the three random-effects models: the estimate of inconsistency  $I^2$  and its corresponding heterogeneity correction  $1/(1-I^2)$ , the estimate of diversity  $D^2$  and its corresponding heterogeneity correction  $1/(1-D^2)$ , and the estimate of between-trial variance,  $\tau^2$ . The estimate of inconsistency is only displayed for the DL model. Note that the estimate of between-trial variance is the same for the DL and BT models (see section 2.1.3). In the lower right corner, there is an option to choose the number of decimal points that all quantities should be displayed with. Click on the drop down window to select the number of decimal points.



**Figure 55** Diversity tab.

## 5. TSA example applications

### 5.1. Datasets

To illustrate the TSA applications, we use data from several published systematic reviews. Some of the analyses and applications presented in this chapter are our own modifications and additions to those that can be found in the original publication.

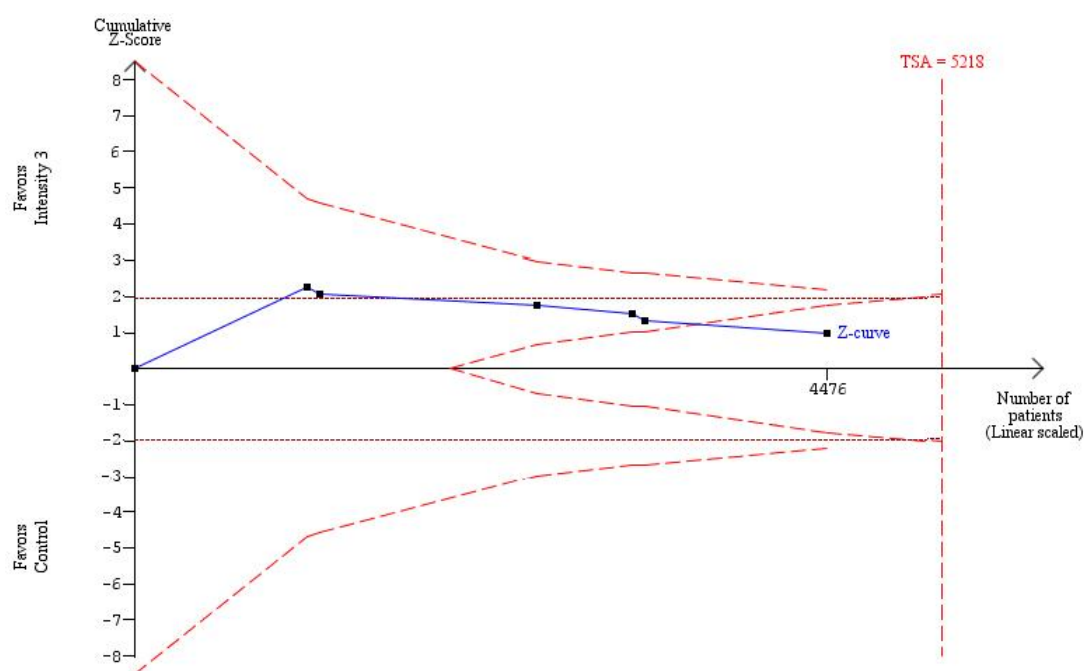
### 5.2. Avoiding false positives

In this example, we used data from a review comparing smoking cessation rates in patients receiving hospital contact plus follow-up for less than 1 month with patients receiving no intervention.<sup>64</sup> In the systematic review, the interventions and length of follow-up differed substantially across the included trials. The authors therefore used the following categorisation of intervention intensity:<sup>64</sup>

1. Single contact in hospital lasting  $\leq 15$  minutes, no follow-up support.
2. One or more contacts in hospital lasting  $>15$  minutes, no follow up support.
3. Any hospital contact plus follow-up  $\leq 1$  month.
4. Any hospital contact plus follow-up  $> 1$  month.

The meta-analysis of intervention intensity 3 included six trials, 4476 patients, and 628 events. The fixed-effect model yielded a pooled relative risk of 1.05 (95% CI 0.91 to 1.21) (the meta-analysis of odds ratios showed a similar result). The estimated inconsistency ( $I^2$ ) was = 9%, and the estimated diversity ( $D^2$ ) was 10%. We performed a retrospective trial sequential analysis, by re-doing a conventional meta-analysis on the accumulating data, each time a new trial was published. The first published trial yielded a relative risk of 1.47 (95% CI 1.05 to 2.05). After the second trial, the pooled relative risk was 1.33 (95% CI 1.02 to 1.75). The meta-analysis comparing intervention intensity category 3 (see above) with control was therefore nominally statistically significant after the first two trials.

We performed TSA on these data. We calculated the information size required to demonstrate or reject a 20% relative benefit increment (smoking cessation being the outcome of benefit). We assumed a 14% event proportion in the control group, which was roughly the median and average control group event proportion. We used a type I error of 5% and a type II error of 20%. We did not correct for heterogeneity. With these settings, we calculated the required information size to 5218 patients. As the number of patients included in the meta-analysis did not exceed the required information size, we also applied futility boundaries to potentially facilitate a firm 'negative' conclusion.



**Figure 56** The required information size to demonstrate or reject a 20% relative increase in benefit on smoking cessation with a control group proportion of 14%, an alpha of 5% and a beta of 20% is 5218 patients (vertical red line). The red dashed lines represent the trial sequential monitoring boundaries and the futility boundaries. The solid blue line is the cumulative Z-curve.

The resulting trial sequential analysis is shown in figure 56. After the first and second trial, the cumulative Z-statistic crossed above 1.96, which corresponds to the nominal threshold for statistical significance, using conventional techniques. From the third trial onwards, the meta-analysis was no longer nominally statistically significant. When the last trial was added, the meta-analysis crossed below the futility boundaries, demonstrating with 80% power

that the effect of an intensity 3 intervention is not larger than a 20% relative increase in smoking cessation. That is, within the set assumptions for confidence and effect size, this intervention is ineffective.

### **5.3. Confirming a positive result**

To illustrate the application of TSA for asserting positive results, we used data from a systematic review comparing off-pump and on-pump coronary artery bypass grafting surgery (CABG).<sup>65</sup>

For this example, the adjusted significance boundaries for the cumulative Z-curve were constructed under the assumption that significance testing may have been performed each time a new trial was added to the meta-analysis. Given the considerable amount of attention that has been given to the off-pump vs on-pump debate over the last decade, this assumption seemed reasonable. We used the monitoring boundaries based on the O'Brien-Fleming type alpha-spending function, which are relatively insensitive to the number of repeated significance tests (see section 2.2.3).

In the considered meta-analysis data sets, there were some years when more than one trial was published. For these years, we have analysed trials in alphabetical order, according to the last name of the first author.

#### **5.3.1. Confirming the 'answer is in'**

To illustrate the application of TSA for asserting 'the answer is in', we used the outcome of atrial fibrillation in this on-pump vs off-pump meta-analysis. Occurrence of atrial fibrillation was reported in 30 trials, including 3634 patients (with two zero-event trials).<sup>65</sup> According to conventional standards for significance testing, off-pump CABG was significantly superior to on-pump CABG in reducing atrial fibrillation (RR 0.69; 95% CI 0.57 to 0.83) (Figure 57). The estimated inconsistency was  $I^2 = 47.3\%$ , and the estimated diversity was  $D^2 = 49.0\%$ .

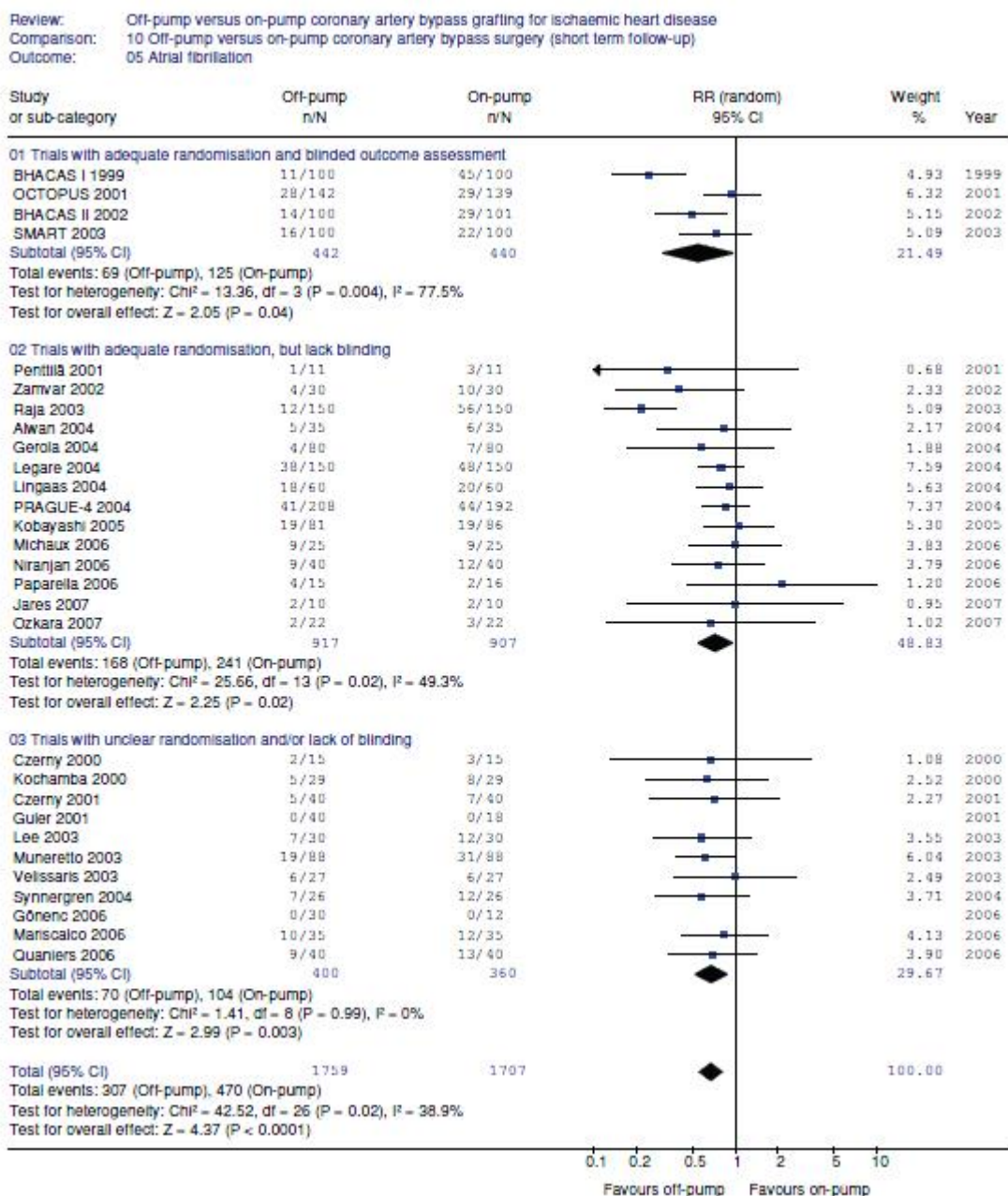


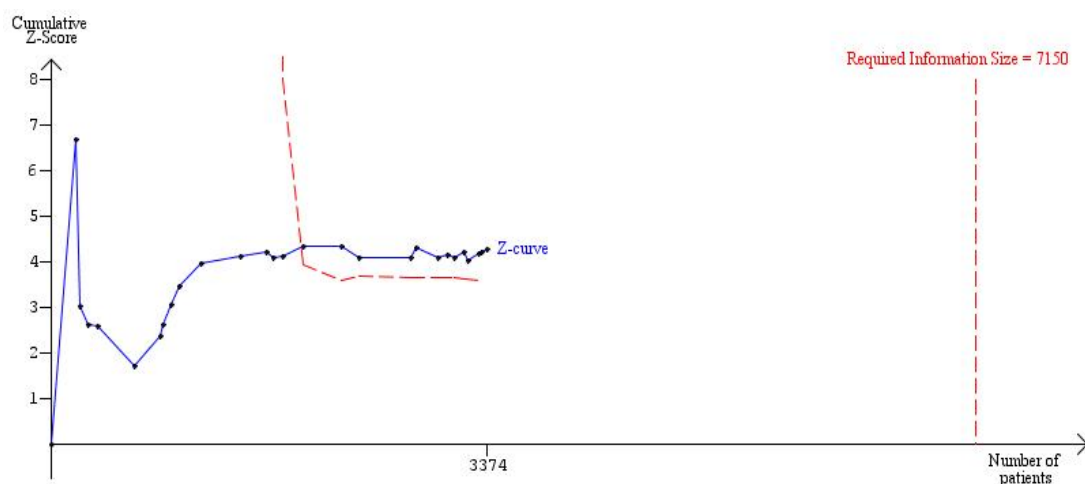
Figure 57 Forest plot of the effect of off-pump vs. on-pump CABG on atrial fibrillation.

In the meta-analysis of trials with low risk of bias (1050 patients), the effect was not significant (0.63, 0.35 to 1.13), the estimated heterogeneity was  $I^2 = 77\%$ , and the estimated diversity was  $D^2 = 79.0\%$ .

### Trial sequential analysis of atrial fibrillation

We calculated two required information sizes for this example. First, we calculated the information size required to demonstrate or reject an *a priori* anticipated intervention effect of a 20% relative risk reduction, alpha of 1%

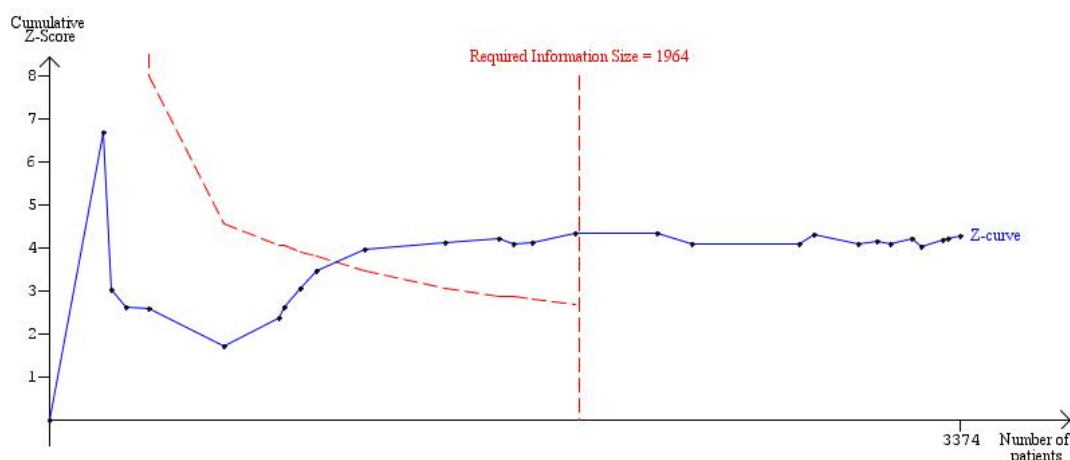
and beta of 10%, which was 7150 patients. The value of 20% anticipated intervention effect was chosen because it was believed to represent a reasonable intervention effect in this clinical situation. Second, we calculated the information size for the meta-analysed estimate of the relative risk reduction from the low-bias-risk trials included in the review (36.9%), which was 1964 patients.



**Figure 58** The heterogeneity-adjusted required information size to demonstrate or reject a 20% relative risk reduction (*a priori* estimate) of atrial fibrillation (with a control group proportion of 27.6%, an alpha of 1%, and a beta of 10%) is 7150 patients (vertical red dashed line). The red dashed inward-sloping line to the left represents the trial sequential monitoring boundaries which are truncated for the first 14 trials. The solid blue line is the cumulative Z-curve.

All information sizes were derived to ensure a maximum type I error of 1%, and a maximum type II error of 10% (i.e., 90% power). All information sizes were heterogeneity adjusted, using the estimate of diversity,  $D^2$ . Both information sizes were derived assuming an event proportion of 27.6% in the on-pump group (median event proportion in this control group).

The cumulative Z-curve crossed the monitoring boundaries constructed from both information size calculations (Figure 58 and 59), thereby confirming that off-pump CABG is superior to on-pump CABG in reducing atrial fibrillation.



**Figure 59** The heterogeneity-adjusted required information size to demonstrate or reject a 36.9% relative risk reduction (low-bias risk trial estimate) of atrial fibrillation (with a control group proportion of 27.6%, an alpha of 1%, and a beta of 10%) is 1964 patients (vertical red dashed line). The red dashed inward-sloping line to the left make up the trial sequential monitoring boundaries which are truncated for the first 4 trials. The solid blue line is the cumulative Z-curve.

### 5.3.2. Avoiding early overestimates

This same example, of atrial fibrillation in CABG, can be used to illustrate how overestimates of effect can happen early in the conventional meta-analytic process. The meta-analysis of atrial fibrillation became statistically significant according to the conventional criterion ( $p < 0.05$ ) after the first trial. All except one of the subsequent P values in the cumulative meta-analysis were also smaller than 0.05. In fact, most subsequent P values were smaller than 0.01. Empirical evidence suggests that pooled effect estimates, even when statistically significant, are unstable when only a limited number of events and patients have been accrued.<sup>4;5;9;29</sup> Insisting that a meta-analysis surpasses its required information size may ensure reliable pooled estimates.<sup>1;2;4;6;19;23</sup>

Table 3 shows the evolution of treatment effects over time, in this example, at the end of each year. The pooled relative risk was grossly overestimated in the first two years and supported by P values smaller than 0.01 (the conventional 99% confidence intervals precluded 1.00). The following three years, the pooled relative risk was overestimated by an absolute risk of at least 10%. In 2003, the meta-analysis crossed the monitoring boundaries from the information size calculation based on the low bias risk estimates,



and in 2004, the meta-analysis surpassed this required information size. In 2004, the meta-analysis also crossed the monitoring boundaries based on a 20% *a priori* relative risk reduction. Both the conventional and adjusted confidence intervals converged between 2002 and 2004.

Year	Total number of			Pooled Effect	99% Confidence Interval	
	Trials	Events	Patients		Conventional	Adjusted
1999	1	55	200	0.24	0.14 to 0.42	0.03 to 7.74
2000	3	74	288	0.39	0.15 to 0.99	0.02 to 7.18
2001	5	143	649	0.57	0.24 to 1.34	0.12 to 2.87
2002	8	204	932	0.52	0.30 to 0.90	0.22 to 1.21
2003 <sup>a</sup>	10	285	1168	0.55	0.37 to 0.81	0.35 to 0.85
2003 <sup>b</sup>	13	391	1722	0.53	0.35 to 0.79	0.34 to 0.83
2004	19	641	2832	0.61	0.46 to 0.82	-
2005	20	679	2999	0.63	0.49 to 0.85	-
2006	25	768	3310	0.67	0.53 to 0.86	-
2007	27	775	3372	0.67	0.53 to 0.86	-

<sup>a</sup> First crossing of the boundaries, <sup>b</sup> End of the year

**Table 3** Shows the evolution of pooled effects (relative risk estimates), conventional and adjusted 99% confidence intervals at the end of each year, with respect to the cumulative number of trials, events, and patients. The adjusted 99% confidence intervals are based on alpha-spending in relation to the required information size (1964 patients), using the relative risk estimate suggested by the trials with low risk of bias.

This example illustrates why pooled estimates based on a relatively small number of events and patients (in this case, less than 100 events and less than 300 patients) should not be trusted. Point estimates from this meta-analysis did not appear to be sufficiently reliable until at least about one hundred events and one thousand patients were accrued. Adjusted confidence intervals serve to guard against spurious inferences at early stages of a meta-analysis, and appropriately converge to resemble conventional confidence intervals as the accrued number of patients approaches the required information size.

Review: Off-pump versus on-pump coronary artery bypass grafting for ischaemic heart disease  
Comparison: 10 Off-pump versus on-pump coronary artery bypass surgery (short term follow-up)  
Outcome: 03 Myocardial infarction

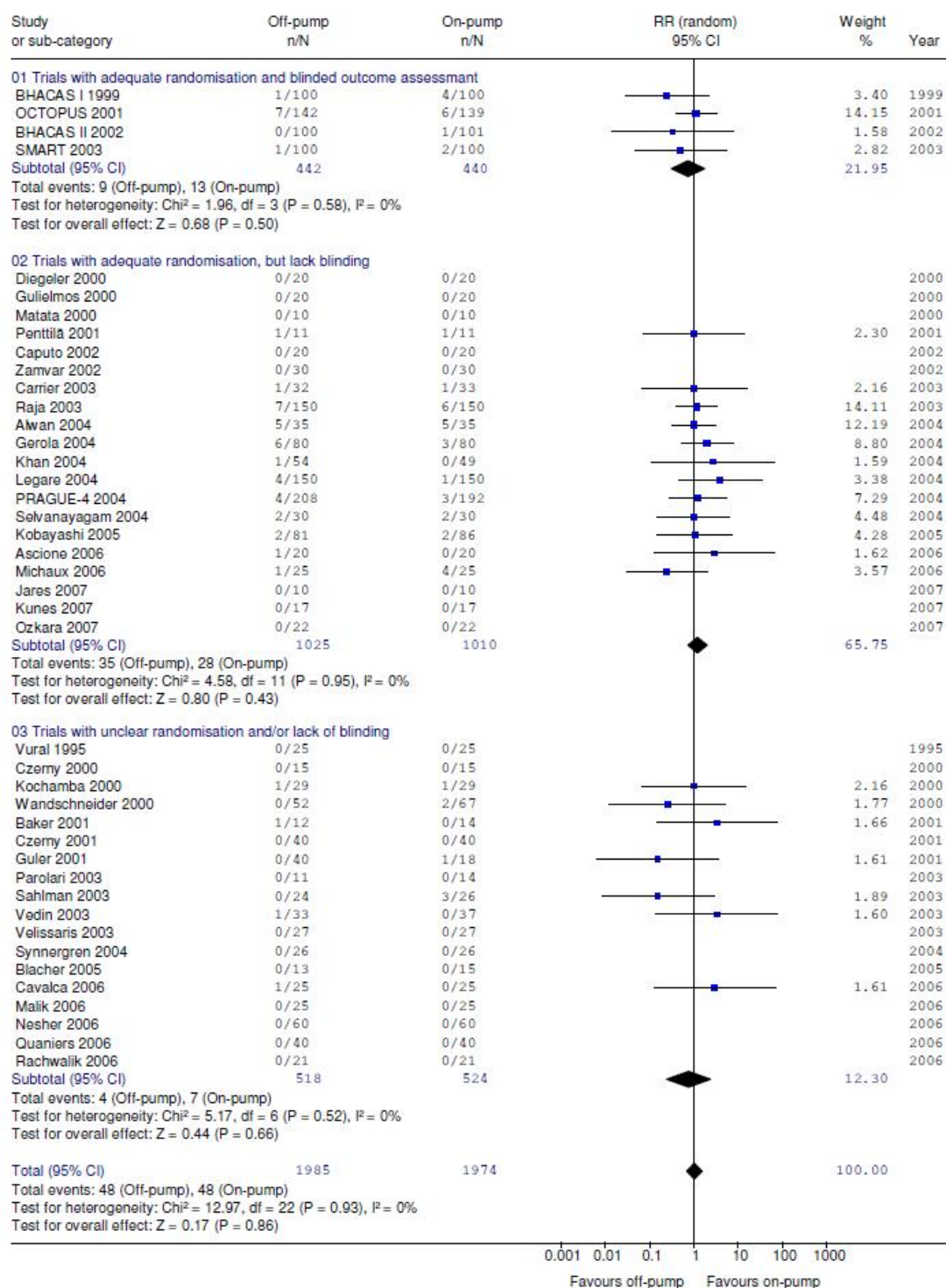
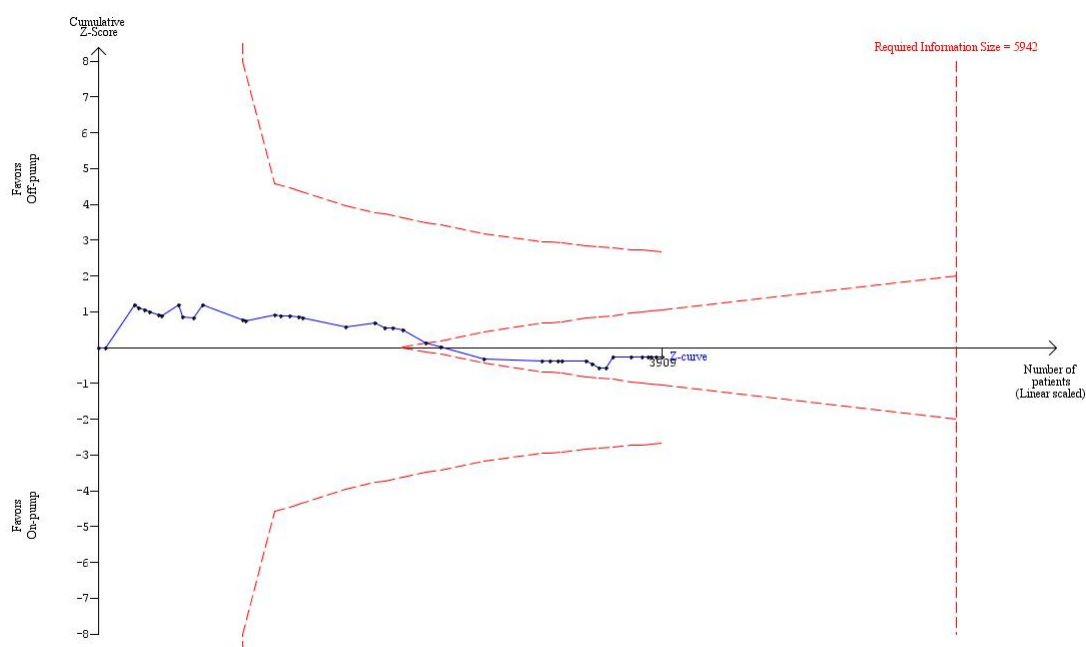


Figure 60 Forest plot of the effect of off-pump vs on-pump CABG on myocardial infarction.

#### 5.4. Testing for futility

The example of the off-pump vs on-pump CABG meta-analysis can also be used to illustrate testing for futility, this time using the outcome of myocardial

infarction (MI). Occurrence of MI was reported in 44 trials including 4303 patients.<sup>65</sup> No significant difference occurred between off-pump vs on-pump surgery (RR 1.06; 95% CI 0.72 to 1.54) (Figure 60) and this result was independent of risk of bias. No statistical heterogeneity was detected ( $I^2 = 0\%$ ). Nineteen trials (909 patients) were zero-event trials. When zero-event trials were continuity corrected, there was also no noticeable change in the results (RR 1.05; 95% CI 0.74 to 1.48).



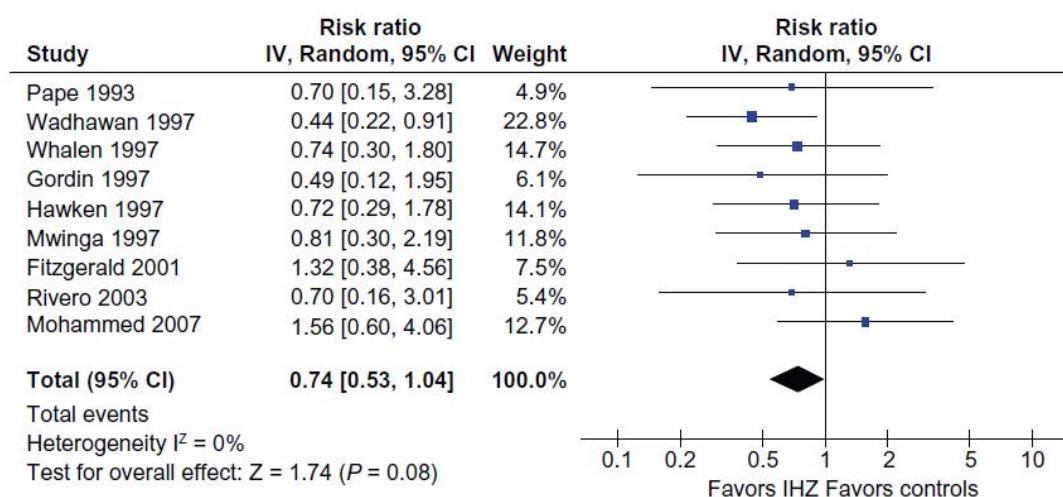
**Figure 61** The heterogeneity-adjusted required information size to demonstrate or reject a 33% relative risk reduction (*a priori* estimate) of myocardial infarction (MI) (with an occurrence of MI in the on-pump group of 3.9%, an alpha of 5%, and a beta of 20%) is 5942 patients (vertical red line). To the left, the red, inward-sloping, dashed lines make up the trial sequential monitoring boundaries. To the right, the red outward sloping dashed lines make up the futility region. The solid blue line is the cumulative Z-curve.

We calculated the information size required to demonstrate or reject an *a priori* anticipated intervention effect of a 33% relative risk reduction. The value of 33% was chosen because it was believed to represent a reasonable intervention effect in this clinical situation. In contrast to the information size calculation for atrial fibrillation, we used a maximum type I error of 5%, and a maximum type II error of 20% (80% power). We used the median proportion of myocardial infarctions in the 'on-pump' groups (excluding the zero-event trials) as our control group event proportion (3.9%). Collectively, these

assumptions yielded a required information size of 5942. The cumulative Z-curve crossed the futility boundaries (Figure 61), and we are therefore able to infer that neither off-pump CABG nor on-pump CABG is more than 33% more effective than the other. This finding, of course, comes with a 20% risk of being a ‘false futile’ finding (the type II error is 20%).

### 5.5. Estimating the sample size of a new clinical trial

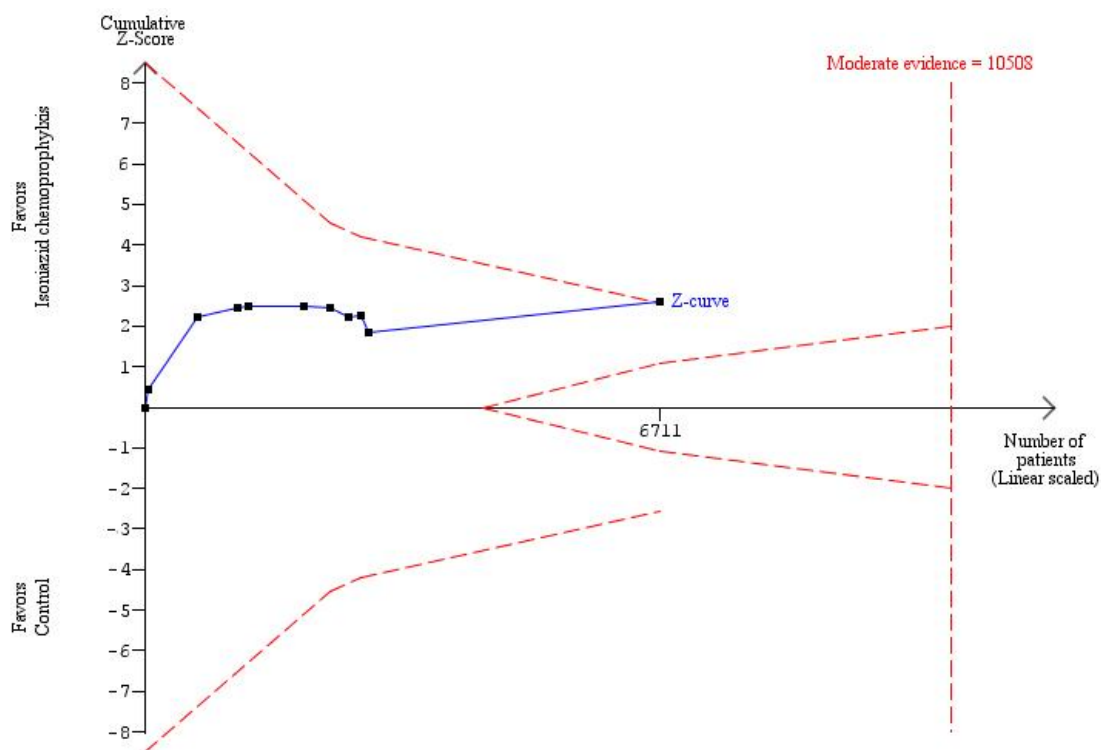
When a meta-analysis has neither crossed the monitoring boundaries nor the futility boundaries, it is possible to approximate how many patients should be randomised in the next trial to make the meta-analysis cross either of the two boundaries. A recent methodology paper illustrated this approach using a meta-analysis of isoniazid chemoprophylaxis for preventing tuberculosis in HIV positive patients.<sup>25</sup> This meta-analysis included nine trials, 2911 patients, and 131 events and yielded a pooled relative risk of 0.74 (95% CI 0.53 to 1.04). The estimated inconsistency and diversity were both 0%.



**Figure 62** Forest plot of the individual trial effects of isoniazid chemoprophylaxis vs. control for preventing tuberculosis in purified protein derivative negative HIV-infected individuals.

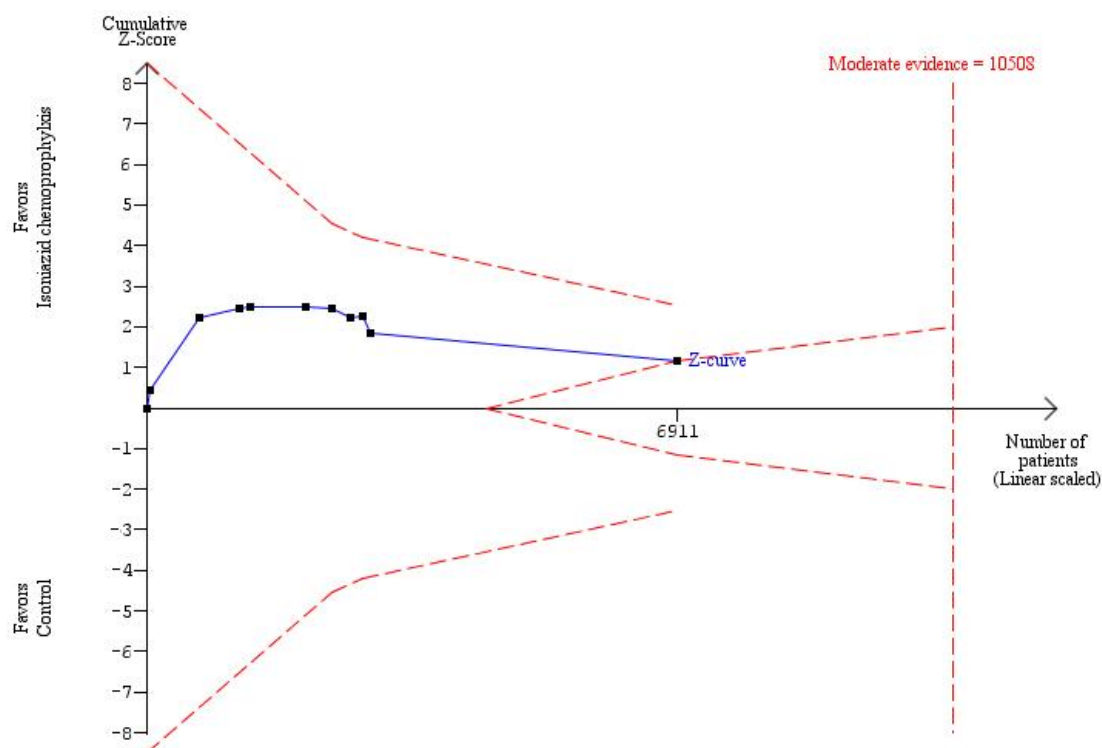
We estimated the required information size for detecting a 25% relative risk reduction in tuberculosis with an alpha = 5% and beta = 20% (80% power). The required information size was based on the assumption of a 5% control group incidence rate (approximately the median rate across trials). We also heterogeneity corrected the required information size assuming 20% diversity ( $D^2$ ). This yielded a required information size of 10,508 patients. Statistical

monitoring boundaries and futility boundaries were subsequently constructed according to the set error levels and the required information size.



**Figure 63** Prospective trial sequential analysis of isoniazid vs control for preventing tuberculosis. To the left, the red inward-sloping dashed lines make up the trial sequential monitoring boundaries. To the right, the outward sloping red dashed lines make up the futility region. The solid blue line is the cumulative Z curve. The last line on the cumulative Z curve represents an imagined trial that makes the meta-analysis conclude that the isoniazid prevents tuberculosis.

To estimate how many patients would needed to be randomised in a future clinical trial to make the meta-analysis conclusive, we approximated the number of patients in an imaginary future trial that would make the cumulative Z curve cross the monitoring boundaries or the futility boundaries. If a future clinical trial were to make the meta-analysis conclusive with a positive result, we assumed that the trials would have the same control group event proportion and intervention effect as hypothesized in the information size considerations. That is, we assumed a trial would have a 5% control group event proportion and yield a 25% relative risk reduction (i.e., the trial would have a 3.75% intervention group event proportion).



**Figure 64** Prospective trial sequential analysis of isoniazid vs control for preventing tuberculosis. To the left, the red inward-sloping dashed lines make up the trial sequential monitoring boundaries. To the right, the red, outward sloping dashed lines make up the futility region. The solid blue line is the cumulative Z curve. The last line on the Z curve represents an imagined trial that makes the meta-analysis conclude that isoniazid is not able to prevent tuberculosis.

If a future clinical trial were to make the meta-analysis conclusive with a futile result, we assumed that the intervention group event proportion would also be 5% (i.e., no effect). We approximated that about 3800 patients (1900 patients in each intervention group) would be required to yield a conclusive positive meta-analysis (Figure 63). About 4000 patients (2000 patients in each intervention group) would be required to yield a conclusive meta-analysis showing futility (Figure 64).

## 5.6. Other published trial sequential analysis applications

The authors of this manual have authored several systematic reviews for which trial sequential analysis was applied to at least one meta-analysis.<sup>14;24;63;65-74</sup> Table 4 provides a brief overview of these publications (ordered by year of publication).

First author	Journal (year)	Meta-analyses
Bangalore <sup>75</sup>	<i>BMJ</i> (2011)	Angiotensin receptor blockers (ARB) vs control for i) non-fatal myocardial infarction ii) all-cause mortality iii) cardiovascular mortality iv) angina pectoris v) stroke vi) heart failure vii) new onset diabetes
Bangalore <sup>76</sup>	<i>Archives of Neurology</i> (2011)	Carotid artery stenting (CAS) vs carotid endarterectomy on i) death, myocardial infarction or stroke ii) periprocedural death or stroke iii) periprocedural stroke
Bangalore <sup>77</sup>	<i>Lancet Oncology</i> (2011)	i) Angiotensin receptor blockers vs. comparison: effect on cancer risk and on cancer-related death ii) Angiotensin converting enzyme inhibitors vs. comparison: effect on cancer risk and on cancer-related death iii) Beta-blockers vs. comparison: effect on cancer risk and on cancer-related death iv) Calcium channel blockers vs. comparison: effect on cancer risk and on cancer-related death v) Diuretics vs. comparison: effect on cancer risk and on cancer-related death
Afshari A <sup>24</sup>	<i>The Cochrane Library</i> (2010)	i) Inhaled nitric oxide vs control for acute respiratory distress syndrome ii) Inhaled nitric oxide vs control for lung injury
Awad T <sup>63</sup>	<i>Hepatology</i> (2010)	Peginterferon alfa-2a vs peginterferon alfa-2b for hepatitis C
Brok J <sup>66</sup>	<i>J Alim Pharm &amp; Ther</i> (2010)	Ribavirin plus interferon vs interferon for hepatitis C
Nielsen N <sup>70</sup>	<i>Int J Cardiol</i> (2010)	Hypothermia vs control after cardiac arrest
Tarnow-Mordi WO <sup>72</sup>	<i>Pediatrics</i> (2010)	i) Probiotics vs control to reduce mortality in newborn ii) Probiotics vs control to reduce necrotizing enterocolitis in newborn
Knorr U <sup>69</sup>	<i>Psychoneuroendocrinology</i> (2010)	Salivary cortisol in depressed patients vs control persons
Bangalore S <sup>14</sup>	<i>The Lancet</i> (2009)	i) Perioperative beta-blockade vs placebo for mortality i) Perioperative beta-blockade vs placebo for myocardial infarction
Brok J <sup>67</sup>	<i>The Cochrane Library</i> (2009)	Ribavirin monotherapy vs placebo for hepatitis C

Whitfield K <sup>73</sup>	<i>The Cochrane Library</i> (2009)	Pentoxifylline vs control for alcoholic hepatitis
Moller CH <sup>65</sup>	<i>Europ Hearj J</i>	i) Off-pump vs on-pump CABG for atrial fibrilation ii) Off-pump vs on-pump CABG for myocardial infarction
Ghandy GY <sup>68</sup>	<i>Mayo Clin Proc</i> (2008)	i) Perioperative insulin infusion vs control for Mortality ii) Perioperative insulin infusion vs control for Morbidity
Rambaldi A <sup>71</sup>	<i>J Alim Pharm &amp; Ther</i> (2008)	Glucocorticosteroids vs control for alcoholic hepatitis
Whitlock R <sup>74</sup>	<i>Europ Heart J</i> (2008)	Prophylactic steroid use vs control for patients undergoing cardiopulmonary bypass
Afshari A <sup>24</sup>	<i>BMJ</i> (2007)	Antithrombin III vs control for reducing cardiac...

---

**Table 4** Overview of published meta-analyses where trial sequential analysis was applied.



## 6. Appendixes

### 6.1. Effect measures for dichotomous and continuous data meta-analysis

The standard errors of the respective effect measures are calculated similarly to the methods used in Review Manager v.5.<sup>27</sup>

For each trial, we denote the number of observed events (e.g., deaths) in the two intervention groups,  $e_A$  and  $e_B$ , and the total number of participants,  $n_A$  and  $n_B$ , in the two intervention groups.

The standard errors for risk differences, relative risks, and odds ratios are calculated using the following formulas:

$$se(RD) = \sqrt{\frac{e_A(1-e_A)}{n_A^3} + \frac{e_B(1-e_B)}{n_B^3}}$$

$$se(RR) = \sqrt{\frac{1}{e_A} + \frac{1}{e_B} - \frac{1}{n_A} - \frac{1}{n_B}}$$

$$se(OR) = \sqrt{\frac{1}{e_A} + \frac{1}{e_B} + \frac{1}{(1-e_A)} + \frac{1}{(1-e_B)}}$$

For a Peto's odds ratio, the standard error is given by:

$$se(OR) = \sqrt{1/v}$$

where

$$v = \frac{(n_A n_B (e_A + e_B) ((1-e_A) + (1-e_B)))}{(n_A + n_B)^2 (n_A + n_B - 1)}$$

## 6.2. Random-effects approaches

### 6.2.1. Formulas for the Biggerstaff-Tweedie method

Let  $f_{DL}(t)$  denote the probability density function of the DL estimate of  $\tau^2$  and let  $F_{DL}(t)$  denote the corresponding cumulative distribution function  $F_{DL}(t)$ . Defining the trial weights as a function of  $t$  by  $w_i(t) = (\sigma_i^2 + t)^{-1}$  and using the obtained distribution of the estimate of  $\tau_{DL}^2$  the so-called frequentist-Bayes estimates of the trial weights can be obtained:

$$\begin{aligned} w_i^*(\tau^2) &= E[w_i^*(\tau_{DL}^2)] \\ &= F(t \cdot 1_{(0,\infty)}(t))w_i(0) + \int_0^\infty w_i^*(t)f_{DL}(t)dt \end{aligned}$$

subsequently yielding summary estimates of the overall population intervention effect:

$$\mu_{BT} = \frac{\sum_i w_i^* Y_i}{\sum_i w_i^*}$$

with variance

$$Var(\mu_{BT}) = \frac{1}{\left(\sum_i w_i^*(\tau_{DL}^2)\right)^2} \left[ \sum_i \left( w_i^*(\tau_{DL}^2) \right) \left( s_i^2 + \tau_{DL}^2 \right) \right]$$

thereby ensuring that the variance of the summary effect estimate is adjusted with regard to the uncertainty associated with estimating the between-trial variance.

## 6.3. Trial sequential analysis

### 6.3.1. Exaggerated type I error due to repeated significance testing

By the laws of basic probability theory, when data is tested twice over time, and when an  $\alpha$  of 5% is used as a threshold for both tests (or a Z value of 1.96), the probability that the two interventions will be declared statistically significant under the null hypothesis is:

$$\begin{aligned}
 \Pr(H_0 \text{ rejected}) &= \Pr(|Z_1| \geq 1.96 \text{ or } |Z_2| \geq 1.96) \\
 &= \Pr(|Z_1| \geq 1.96) \cdot \Pr(|Z_2| \geq 1.96 \mid |Z_1| < 1.96) \\
 &= (1 - \Pr(|Z_1| < 1.96)) \cdot (1 - \Pr(|Z_2| < 1.96 \mid |Z_1| < 1.96)) \\
 &= 1 - \Pr(|Z_1| < 1.96) - \Pr(|Z_2| < 1.96 \mid |Z_1| < 1.96) + \Pr(|Z_1| < 1.96) \cdot \Pr(|Z_2| < 1.96 \mid |Z_1| < 1.96) \\
 &= 1 - \Pr(|Z_1| < 1.96) - \Pr(|Z_2| < 1.96 \mid |Z_1| < 1.96) + \Pr(|Z_1| < 1.96 \text{ or } |Z_2| < 1.96) \\
 &= 0.05 - \Pr(|Z_2| < 1.96 \mid |Z_1| < 1.96) + \Pr(|Z_1| < 1.96 \text{ or } |Z_2| < 1.96) \\
 &> 0.05
 \end{aligned}$$

Where the inequality is apparent from the fact that

$$\Pr(|Z_2| < 1.96 \text{ or } |Z_1| < 1.96) > \Pr(|Z_2| < 1.96 \mid |Z_1| < 1.96)$$

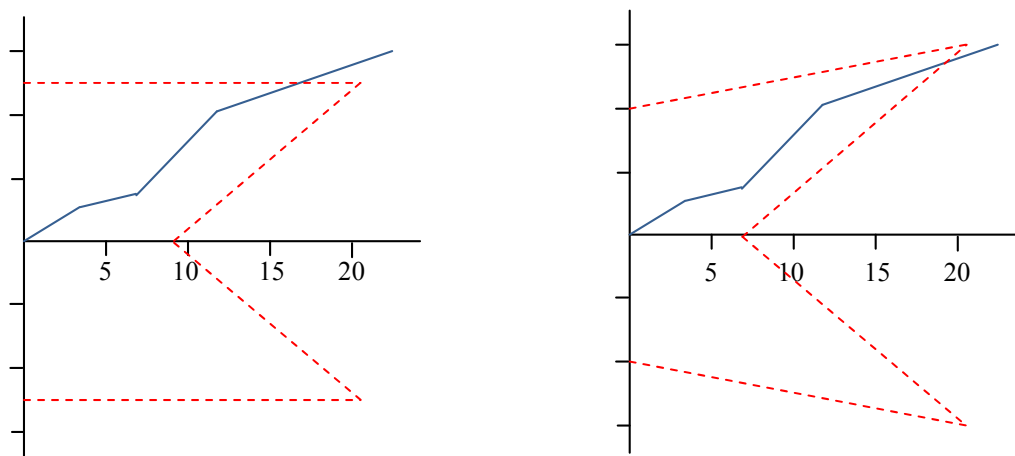
The above is easily generalisable for any value of  $\alpha$  and for any number of repeated significance tests.

### **6.3.2. Alternative methods not implemented in the TSA software**

A wide range of methods are available for repeated significance testing in randomised clinical trials – some of which may also find application in meta-analysis.<sup>30</sup> The approaches implemented in the TSA software are all approaches constructed around monitoring of the standardized Z-statistic (or at least an adjustment hereof). Other sequential approaches which have received some attention in the context of meta-analysis are constructed to monitor other statistics.

One approach that has recently received some attention is the sequential analysis (monitoring) of *efficient scores* or the *likelihood score statistic* for the meta-analysed effect.<sup>78-81</sup> In the standard meta-analysis setting the *efficient score* for each trial is simply the estimated trial treatment effect multiplied by its variance, and the efficient score for a meta-analysis is the sum of trial efficient scores. In sequential analysis of efficient scores, information is measured as statistical information (i.e., Fischer information). The efficient score is plotted (y-axis) against the statistical information (x-axis), and monitored with some boundaries. Just as with the alpha-spending and beta-

spending based boundaries, the sequential method for monitoring efficient scores produce superiority, inferiority, and futility boundaries. Examples of such boundaries are illustrated in figure 65 below.



**Figure 65** Illustration of two types of monitoring boundaries from sequential meta-analysis of efficient scores. The left graph illustrates what would correspond to an O'Brien-Fleming alpha-spending significance boundaries and O'Brien-Fleming beta-spending futility boundaries. The right graph illustrates what corresponds to what is typically known and Whitehead's triangular boundaries. The latter is designed to minimize total risk of statistical error (i.e., type I and type II error together).

Just like different  $\alpha$ -spending functions yield different types of adjusted significance boundaries, the triangular test can be used to construct different types of boundaries (and similarly for beta-spending functions and futility boundaries).<sup>45;80</sup> For example, a special case of the triangular test yields boundaries that are equivalent to the O'Brien-Fleming boundaries when accumulating statistical information (left graph in figure 65).

The O'Brien-Fleming type efficient score sequential boundaries were recently explored empirically and through simulation.<sup>80</sup> A study by van der Tweel and Bollen compared O'Brien-Fleming significance boundaries (the ones implemented in the TSA software) to the O'Brien-Fleming type efficient score sequential boundaries in six meta-analysis.<sup>80</sup> These six meta-analyses were the ones initially (and randomly) selected as illustrative examples in the methods paper proposing the information size heterogeneity correction for trial sequential analysis which is described in section 2.2.1. of this manual.<sup>6</sup>

Tweel and Bollen found that the two methods were identical in testing for significance. A simulation study by Higgins et al investigated the type I error and adjusted confidence interval coverage associated with the O'Brien-Fleming type efficient score sequential boundaries under a number of random-effects model approaches.<sup>79</sup> They found that under this design the conventional DerSimonian-Laird random-effects model and the Biggerstaff-Tweedie approach did not yield satisfactory results, but a semi-Bayesian approach utilizing an informative Gamma distribution on the between-trial variance did. Another example of the efficient score sequential boundaries is the triangular test proposed by Whitehead.<sup>45;81</sup> The boundaries produced from this method are illustrated in the right graph of figure 65. The triangular test boundaries are statistically constructed to yield the minimum possible risk of committing an error (either a type I error or type II error).<sup>30;45</sup> This emphasis - on minimising both types of error - skews this technique towards favouring total risk of error over risk of type I error. In the context of medical research, conventional theory does not support this balance; prevention of alpha error has always been considered more important.

The performance of the Whitehead triangular test applied in meta-analysis has been explored in a simulation study, where the method was found to exhibit poor control of the maximum type I error in heterogeneous meta-analyses.<sup>81</sup> The results of this study suggested that the more heterogeneous a meta-analysis data set is, the worse the triangular test exhibits control of the type I error.<sup>81</sup> To date, the literature contains one example of the Whitehead triangular test being applied to meta-analysis comparing death or chronic lung disease after high-frequency ventilation with conventional mechanical ventilation in the treatment of preterm newborns.<sup>78</sup> In this example, the meta-analysis demonstrated no difference between the two interventions as the cumulative score statistic crossed the *futility boundaries*.<sup>78</sup>

*Stochastic curtailment* is another method for controlling the risk of false positives and false negatives.<sup>1;2</sup> When applied to meta-analysis, this method concentrates on predicting what the outcome will be once a meta-analysis surpasses its required information size.<sup>1;2</sup> More specifically, stochastic

curtailment is a method for calculating the likelihood that the current trend of the data will reverse before surpassing the required information size. When the probability of such a reversal is sufficiently small, a meta-analysis may be considered conclusive. Two conditional probabilities can be calculated. First, if the current trend in the data is suggesting that the experimental intervention is effective, stochastic curtailment may be used to calculate the probability of rejecting the null hypothesis when the meta-analysis surpassed the required information size. If this conditional probability is sufficiently high, the meta-analysis can be considered conclusive. Similarly, if the current data is suggesting a lack of trend, stochastic curtailment can be utilised to calculate the probability of failing to reject the null hypothesis once the meta-analysis surpasses its required information size. Again, if this conditional probability is sufficiently high, the meta-analysis can be considered conclusive. Stochastic curtailment may be a valuable tool to assist decision making from formal significance testing methods. However, because most meta-analyses are subject to some time-trend bias, the conditional probability of a trend reversal is very likely to be biased as well.

## 7. List of abbreviations and statistical notation

The following chapter provides a guide to the abbreviations, notation, and terminology used in this manual. In some cases, these definitions will vary from other sources. Our intention is to provide the reader with a guide for how these terms were used in this manual.

### 7.1. General abbreviations

AF	- Adjustment Factor
BT	- Biggersaff-Tweedie
CI	- Confidence Interval
$D^2$	- Diversity
DL	- DerSimonian-Laird
$I^2$	- Inconsistency
IF	- Information Fraction
IS	- Information Size
JRE	- Java Runtime Environment
MD	- Mean Difference
OIS	- Optimal Information Size
OR	- Odds Ratio
RCT	- Randomised Controlled Trial
RD	- Risk Difference
RR	- Relative Risk
RRR	- Relative Risk Reduction
SJ	- Sidik-Jonkman
SMD	- Standardised Mean Difference
TSA	- Trial Sequential Analysis

### 7.2. Statistical notation

#### 7.2.1. Lower case letter symbols

$c$	– The statistical significance threshold with respect $ Z $
$c_i$	– The adjusted threshold for $Z_i$ under repeated testing
$e_x$	– The number of events in intervention group X

$f_{DL}(t)$	– The probability distribution for the DerSimonian-Laird estimator
$k$	– The number of trials in a meta-analysis
$m_X$	– The mean response in intervention group X
$n_X$	– The number of patients in intervention group X
$sd_X$	– The standard deviation in intervention group X
$v$	– Variance estimate
$v_F$	– The variance in a fixed-effect model
$v_R$	– The variance in a random-effects model
$w_i$	– The weight assigned to the $i$ -th trial in a fixed-effect model
$w_i^*$	– The weight assigned to the $i$ -th trial in a random-effects model
$w_i(t)$	– The $i$ -th trial weight as a function of the between-trial variance

### 7.2.2. Upper case letter symbols

$AF$	– The heterogeneity adjustment factor
$C$	– The sum of the continuity corrections for two groups
$CF_X$	– The continuity correction for intervention group X
$D^2$	– The diversity measure used to quantify heterogeneity
$E(X)$	– The expectation of X
$H$	– A conceptual measure of $D^2$
$H_0$	– The null hypothesis
$I^2$	– The inconsistency measure used to quantify heterogeneity
$I_j$	– The cumulative statistical information after the $j$ -th
$IF_i$	– The cumulative information fraction after the $i$ -th trial
$IS_{Patients}$	– The required number of patients in a meta-analysis
$IS_{Events}$	– The required number of events in a meta-analysis
$IS_{Statistical}$	– The required statistical information in a meta-analysis
$IS_{Fixed}$	– The required information size for a fixed-effect model
$IS_{Random}$	– The required information size for a random-effects model
$OR_i$	– The odds ratio estimate of the $i$ -th trial
$P$	– The test P-value derived from Z
$P_X$	– The event rate in intervention group X
$P^*$	– The average event rates of the two treatment groups
$Pr(X)$	– The probability that some event X occurs
$Pr(X Y)$	– The probability that some event X given the event Y occurred



$Q$	– The Cochran homogeneity test statistic
$R$	– The randomisation ratio
$RD_i$	– The risk difference estimate of the $i$ -th trial
$RR_i$	– The relative risk estimate of the $i$ -th trial
$S_r$	– The sum of trial weights to the $r$ -th power
$SE(X)$	– The standard error of $X$
$Var(X)$	– The variance of $X$
$Z$	– The test statistic for whether there exists an intervention effect
$Z_i$	– The $Z$ -value from the meta-analysis including the first $i$ trials
$Z_{1-\alpha/2}$	– The $(1-\alpha/2)$ -th percentile of the standard normal distribution
$Z_{1-\beta}$	– The $(1-\beta)$ -th percentile of the standard normal distribution
$Y_i$	– The observed intervention effect in the $i$ -th trial

### 7.2.3. Greek letter symbols

$\alpha$	– The maximum risk of type I error
$\alpha(t)$	– The cumulative type I error risk as a function of time
$\beta$	– The maximum risk of type II error
$\beta(t)$	– The cumulative type II error risk as a function of time
$\delta$	– The <i>a priori</i> estimate of an anticipated intervention effect
$\delta_F$	– The anticipated intervention effect in a fixed-effect model
$\delta_R$	– The anticipated intervention effect in a random-effects model
$\lambda$	– A constant to ensure control of $\alpha$ when penalising $Z$
$\mu_i$	– The underlying ‘true’ intervention effect of the $i$ -th trial
$\mu$	– The overall ‘true’ intervention effect
$\hat{\mu}$	– The pooled intervention effect
$\sigma^2$	– The variance of $\delta$
$\sigma_i^2$	– The variance of $Y_i$
$\tau^2$	– The between-trial variance
$\tau_{DL}^2$	– The DerSimonian-Laird estimate for the between-trial variance
$\tau_{SJ}^2$	– The Sidik-Jonkman estimate for the between-trial variance
$\hat{\theta}$	– The pooled odds ratio of excluding zero-event trials
$\Phi$	– The cumulative standard normal probability distribution function

#### Reference List

- (1) Pogue J, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clinical Trials* 1997;18:580-593.
- (2) Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet* 1998;351:47-52.
- (3) Sterne JA, Davey SG. Sifting the evidence - what's wrong with significance tests? *British Medical Journal* 2001;322:226-231.
- (4) Thorlund K, Devereaux PJ, Wetterslev J et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *International Journal of Epidemiology* 2009;38:276-286.
- (5) Trikalinos TA, Churchill R, Ferri M et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *Journal of Clinical Epidemiology* 2004;57:1124-1130.
- (6) Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology* 2008;61:64-75.
- (7) Hu M, Cappelleri J, Lan KK. Applying the law of the iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clinical Trials* 2007;4:329-340.
- (8) Lan KK, Hu M, Cappelleri J. Applying the law of the iterated logarithm to cumulative meta-analysis of a continuous endpoint. *Statistica Sinica* 2003;13:1135-1145.
- (9) Ioannidis J, Lau J. Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. *Proc Natl Acad Sci U S A* 2001;98:831-836.
- (10) Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. *J Clin Epidemiol* 2009;62:825-830.
- (11) Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *Journal of Clinical Epidemiology* 2008;61:763-769.
- (12) Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive - Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. *International Journal of Epidemiology* 2009;38:287-298.
- (13) Higgins JPT, Green S. *Cochrane Handbook for systematic reviews of interventions, version 5.0.0*. John Wiley & Sons, 2009.
- (14) Bangalore S, Wetterslev J, Pranesh S, Sawhney S, Gluud C, Messerli FH. Perioperative beta blockers in patients having non-cardiac surgery: a meta-analysis. *Lancet* 2008;372:1962-1976.
- (15) Devereaux PJ, Beattie WS, Choi PT et al. How strong is the evidence for the use of perioperative beta blockers in non-cardiac surgery? Systematic review and meta-analysis of randomised controlled trials. *BMJ* 2005;331:313-321.
- (16) DeMets D, Lan KK. Interim analysis: the alpha spending function approach. *Statistics in Medicine* 1994;12:1341-1352.

- (17) Lan KK, DeMets D. Discrete sequential monitoring boundaries for clinical trials. *Biometrika* 1983;659-663.
- (18) Pocock S. Intermittent analyses for randomized clinical trials: the group sequential approach. *Biometrics* 1982;38:153-162.
- (19) GRADE Working Group. Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. *Allergy* 2009;64:669-677.
- (20) Guyatt G, Mills E. In the era of systematic reviews, does the size of an individual trial still matter. *PLoS Medicine* 2008;5:e4.
- (21) Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. *BMC Medical Research Methodology* 2009;9.
- (22) Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001;357:1191-1194.
- (23) Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in a random-effects meta-analysis. *BMC Medical Research Methodology* 2009;9.
- (24) Afshari A, Wetterslev J, Brok J, Moller A. Antithrombin III in critically ill patients: systematic review with meta-analysis and trial sequential analysis. *BMJ* 2007;335:1248-1251.
- (25) Thorlund K, Anema A, Mills E. Interpreting meta-analysis according to the adequacy of sample size. An example using isoniazid chemoprophylaxis for tuberculosis in purified protein derivative negative HIV-infected individuals. *Clinical Epidemiology* 2010;2:57-66.
- (26) Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. *Lancet* 2002;359:519.
- (27) Review Manager (RevMan) [Computer program]. Version 5.1. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2011.
- (28) Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. *PLoS Medicine* 2010;4:e78.
- (29) Gehr B, Weiss C, Porzolt F. The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Medical Research Methodology* 2006;6:25.
- (30) Jennison C, Turnbull B. *Group sequential methods with applications to clinical trials*. Chapman&Hall/CRC Press, 2000.
- (31) Berkey C, Mosteller F, Lau J. Uncertainty of the time of first significance in random-effects cumulative meta-analysis. *Controlled Clinical Trials* 1996;17:357-371.
- (32) DerSimonian L, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986;7:177-188.
- (33) Sidik K, Jonkman J. Simple heterogeneity variance estimation for meta-analysis. *Journal of Royal Statistical Society(C)* 2005;54:367-384.
- (34) Sidik K, Jonkman J. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine* 2007;26:1964-1981.
- (35) Brockwell S, Gordon I. A comparison of statistical methods for meta-analysis. *Statistics in Medicine* 2001;20:825-840.

- (36) Brockwell S, Gordon IR. A simple method for inference on an overall effect in meta-analysis. *Stat Med* 2007;26:4531-4543.
- (37) Sanchez-Meca J, Martin-Martinez, F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods* 2008;13:31-48.
- (38) Sidik K, Jonkman J. Robust variance estimation for random-effects meta-analysis. *Comp Stat Data An* 2006;50:3681-3701.
- (39) Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educa Behav Stat* 2005;30:261-293.
- (40) Makambi KH. The effect of the heterogeneity variance estimator on some tests of efficacy. *J Biopharm Stat* 2004;14:439-449.
- (41) Biggerstaff B, Tweedie R. Incorporating variability in estimates of heterogeneity in the random-effects model in meta-analysis. *Statistics in Medicine* 2009;16:753-768.
- (42) Sweeting M, Sutton A, Lambert P. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis. *Statistics in Medicine* 2004;23:1351-1375.
- (43) Armitage P. Sequential analysis in therapeutic trials. *Annual Review of Medicine* 1969;20:425-430.
- (44) Pocock S. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977;191-199.
- (45) Whitehead J. *The design and analysis of sequential clinical trials*. John Wiley & Sons, 2000.
- (46) Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* 1991;10:1665-1677.
- (47) Higgins JPT, Thompson S. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002;21:1539-1558.
- (48) Federov V, Jones B. The design of multicentre trials. *Statistical Methods in Medical Research* 2005;14:205-248.
- (49) Ioannidis J, Patsopoulos N, Evangelou E. Uncertainty in heterogeneity estimates in meta-analysis. *British Medical Journal* 2007;335:914-916.
- (50) Jackson D. The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine* 2006;25:2911-2921.
- (51) Chan AW, Hrobjartsson A, Haahr M, Gotzsche P, Altman D. Empirical evidence for selective reporting of outcomes in randomized trials. comparison of protocols to published articles. *Journal of American Medical Association* 2004;291:2457-2465.
- (52) Chan AW, Altman D. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal* 2004;171:735-740.
- (53) Dwan K, Altman D, Arnaiz J et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS Medicine* 2008;3:e3081.
- (54) Hrobjartsson A, Chan AW, Haahr M, Gotzsche P, Altman D. Selective reporting of positive outcomes in randomised trials - secondary publication. A comparison of protocols with published reports. *Ugeskr Laeger* 2005;167:3189-3191.

- (55) Kjaergaard L, Villumsen J, Gluud C. Reported methodological quality and discrepancies between small and large randomized trials in meta-analyses. *Annals of Internal Medicine* 2001;135:982-989.
- (56) Moher D, Pham B, Jones A et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;352:609-613.
- (57) Schulz K, Chalmers I, Hayes R, Altman D. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of American Medical Association* 1995;273:408-412.
- (58) Wood L, Egger M, Gluud LL et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *British Medical Journal* 2008;336:601-605.
- (59) Guyatt GH, Oxman AD, Vist GE et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-926.
- (60) Reboussin DM, DeMets DL, Kyungmann K, Lan KKG. Programs for Computing Group Sequential Boundaries Using the Lan-DeMets Method, v.2. 2009.
- (61) O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549-556.
- (62) Pocock SJ. When to stop a clinical trial. *British Medical Journal* 1992;305:235-240.
- (63) Awad T, Thorlund K, Hauser G, Stimac D, Mabrouk M, Gluud C. Peginterferon alpha-2a is associated with higher sustained virological response than peginterferon alfa-2b in chronic hepatitis C: Systematic review of randomized trials. *Hepatology* 2010;51:1176-1184.
- (64) Rigotti N, Munafo MR, Stead LF. Interventions for smoking cessation in hospitalised patients. *Cochrane Database of Systematic Reviews* 2007.
- (65) Moller CH, Penninga L, Wetterslev J, Steinbruchel DA, Gluud C. Clinical outcomes in randomized trials of off- vs. on-pump coronary artery bypass surgery: systematic review with meta-analyses and trial sequential analyses. *Eur Heart J* 2008;29:2601-2616.
- (66) Brok J, Gluud LL, Gluud C. Meta-analysis: ribavirin plus interferon versus interferon monotherapy for chronic hepatitis C - an updated Cochrane review. *Alimentary Pharmacology and Therapeutics* 2010;32:840-850.
- (67) Brok J, Gluud C. Ribavirin monotherapy for hepatitis C. *Cochrane Database of Systematic Reviews* 2010;Issue 1.
- (68) Ghandi GYMM, Flynn DN, et al. Effect of perioperative insulin infusion on surgical morbidity and mortality: systematic review and meta-analysis of randomized trials. *Mayo Clinique Proceedings* 2008;83:418-430.
- (69) Knorr U, Vinberg M, Kessing LV, Wetterslev J. Salivary cortisol in depressed patients versus control persons: a systematic review and meta-analysis. *Psychoneuroendocrinology* 2010; 35(9):1275-86.
- (70) Nielsen N, Friberg H, Gluud C, Herlitz J, Wetterslev J. Hypothermia after cardiac arrest should be further evaluated - a systematic review of randomised trials with meta-analysis and trial sequential analysis. *International Journal of Cardiology* 2010.
- (71) Rambaldi A, Saconato HH, Christensen E, Thorlund K, Wetterslev J, Gluud C. Systematic review of glucocorticosteroids for alcoholic hepatitis - a Cochrane Hepato-Biliary Group systematic review with meta-analysis and trial sequential analysis of randomised clinical trials. *Alimentary Pharmacology and Therapeutics* 2008;27:1167-1178.

- (72) Tarnow-Mordi WO, Wilkinson D, Trivedi D, Brok J. Probiotics reduce all-cause mortality and necrotizing enterocolitis: it is time to change practice. *Pediatrics* 2010;125:1068-1070.
- (73) Whitfield K, Rambaldi A, Wetterslev J, Gluud C. Pentoxifylline for alcoholic hepatitis C. *Cochrane Database of Systematic Reviews* 2009.
- (74) Whitlock R, Chan S, Devereaux PJ. Clinical benefit of steroid use in patients undergoing cardiopulmonary bypass: a meta-analysis of randomized trials. *European Heart Journal* 2008;29:2592-2600.
- (75) Bangalore S, Kumar S, Wetterslev J, Messerli FH. Angiotensin receptor blockers and risk of myocardial infarction: meta-analyses and trial sequential analyses of 147 020 patients from randomised trials. *BMJ* 2011;342:d2234.
- (76) Bangalore S, Kumar S, Wetterslev J et al. Carotid artery stenting vs carotid endarterectomy: meta-analysis and diversity-adjusted trial sequential analysis of randomized trials. *Arch Neurol* 2011;68:172-184.
- (77) Bangalore S, Kumar S, Kjeldsen SE et al. Antihypertensive drugs and risk of cancer: network meta-analyses and trial sequential analyses of 324,168 participants from randomised trials. *Lancet Oncol* 2011;12:65-82.
- (78) Bollen C, Uiterwaal C, Vught A, van der Tweel I. Sequential meta-analysis of past clinical trials to determine the use of a new trial. *Epidemiology* 2006;17:644-649.
- (79) Higgins JPT, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. *Stat Med* 2011;30:903-921.
- (80) van der Tweel I, Bollen C. Sequential meta-analysis: an efficient decision making tool. *Clinical Trials* 2010;7:136-146.
- (81) Whitehead A. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in Medicine* 1997;16:2901-2913.