Original article

Research Synthesis Methods

Received 24 January 2015, Revised 03 July 2015, Accepted 27 August 2015 Published online 1 December 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1185

Sequential biases in accumulating evidence

Elena Kulinskaya,[a]* Richard Huggins[b] and Samson Henry Dogo[a]

Whilst it is common in clinical trials to use the results of tests at one phase to decide whether to continue to the next phase and to subsequently design the next phase, we show that this can lead to biased results in evidence synthesis. Two new kinds of bias associated with accumulating evidence, termed 'sequential decision bias' and 'sequential design bias', are identified. Both kinds of bias are the result of making decisions on the usefulness of a new study, or its design, based on the previous studies. Sequential decision bias is determined by the correlation between the value of the current estimated effect and the probability of conducting an additional study. Sequential design bias arises from using the estimated value instead of the clinically relevant value of an effect in sample size calculations. We considered both the fixed-effect and the random-effects models of meta-analysis and demonstrated analytically and by simulations that in both settings the problems due to sequential biases are apparent. According to our simulations, the sequential biases increase with increased heterogeneity. Minimisation of sequential biases arises as a new and important research area necessary for successful evidence-based approaches to the development of science. © 2015 The Authors. *Research Synthesis Methods* Published by John Wiley & Sons Ltd.

Keywords: accumulating evidence; cumulative meta-analysis; sequential meta-analysis; sequential bias

For with much wisdom comes much sorrow; the more knowledge, the more grief.

Ecclesiastes 1:18

1. Introduction

The idea that the results of previous meta-analyses should be used for design of new trials is widely recognised. For example, the UK Medical Research Council requires a comprehensive review of existing evidence before funding trials (Glasziou *et al*., 2006). The guidelines of several medical journals including the Journal of American Medical Association and the Lancet state that all reports of clinical trials must include a summary of previous research findings and explain how the new trial affects this summary with direct reference to existing meta-analysis (Goudie *et al*., 2010).

There are two ways of using prior analyses to inform further research. The first is using prior information in making the decision to conduct a new trial (sequential decision). The second is using prior analyses and systematic reviews to design the next trial (sequential design). That is, both the decision to conduct an experiment and the subsequent design of this experiment may depend on the results of previous experiments, and after the new experiment is conducted, the results are combined in an updated meta-analysis.

Sequential and cumulative meta-analyses are established techniques in both fixed-effects and random-effects meta-analyses. See Whitehead (1997), Higgins *et al*. (2011), van der Tweel and Bollen (2010) to name a very few. Often, in a sequential analysis, when the results seem promising, after each trial, the only decision is whether or not to add the next trial in a sequence of independent trials. Whilst not advocating the approach and remarking on its inherent flaws, van der Tweel and Bollen (2010) noted that

[a] *School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK*

[b] *Department of Mathematics and Statistics, University of Melbourne, Melbourne, Australia*

**Correspondence to: Elena Kulinskaya, School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK. E-mail: e.kulinskaya@uea.ac.uk*

> "The usual approach is to repeatedly test the null hypothesis of equal effectiveness of two treatments on the cumulative data. If the test result is not statistically significant, a new trial is added and the test is repeated".

But the sequence of trials may also be stopped for futility, and a systematic review would then reach the conclusion that a new trial is unnecessary (Goudie *et al*., 2010).

The setting explored here differs from the standard sequential meta-analysis setting in that after K studies are accumulated and their results meta-analysed, a meta-analyst has an active role in the decision-making and the design of the subsequent, $(K+1)$-st study, aiming at the definitive meta-analysis of the $(K+1)$ studies. No direct involvement beyond this planned study is assumed. We examine the effect of prior knowledge on decision-making in Section 2, where we show that if the probability of proceeding to a new trial is correlated with the current estimate of the effect size, the new combined estimator from a meta-analysis will be biased (sequential decision bias).

Systematic reviews are often used to inform study design (Cooper *et al*., 2005). The sample size of the new study is calculated to yield a given type-I error rate, typically 5%, and power, typically 80%, to detect a given effect size. 'Researchers use the review to …estimate sample sizes' (Mulrow, 1994). Or,

> "In this context, we would wish to have answers to questions such as how a new study will influence the result of the meta-analysis or how much more information might be needed to make the meta-analysis conclusive (Roloff *et al*., 2013)".

'A new trial may, thus, be designed so as to achieve a decisive result when it is added to an existing meta-analysis.' (Goudie *et al*., 2010). We demonstrate in Section 3 that if the previous experiment is used to determine the effect size used in the sample size calculations in this fashion, then the combined estimator is biased. To illustrate these effects, we consider both fixed-effect and random-effects meta-analyses analytically and by simulation, and in both settings, the problems become apparent. In Section 4, we illustrate the arising biases using as an example the systematic review by Johnson (1993). Similar biases appear and are well understood in group-sequential and adaptive clinical trials. In this area, the existence of sequential bias is widely recognised and the means of adjustment for this bias have been developed (Denne, 2000; Emerson & Fleming, 1990; Kirby *et al*., 2012; Li *et al*., 2002; Whitehead, 1986). Discussion is given in Section 5 where, among other issues, we discuss similarities and differences between our setting and that of drug development. Proofs of some technical results, description of our simulations and additional figures depicting the simulation results are given in the Supporting Information.

2. Sequential decision bias

2.1. Bias

To illustrate sequential decision bias, we consider the following simple situation. Suppose there is a study that had estimated the effect of interest, θ, by $\hat{\theta}_1$ and its variance by s_1^2. A researcher is considering the usefulness of running another study. Suppose that the probability of running this new study $p_{(1)} = p\left(\hat{\theta}_1, s_1^2, \theta_0\right)$ is a function of the estimated effect, its precision and the effect of clinical interest θ_0 that is the same in both studies. In general, we denote by $\hat{\theta}_i$ and s_i^2 the effect and the estimated variance from the ith study, $i \geq 1$. We also adopt sequential notation using bracketed subscripts, so that $\hat{\theta}_{(i)}$ is the meta-analytically combined effect from the first i studies and $s_{(i)}^2$ is its estimated variance. For the first study, $\hat{\theta}_1 = \hat{\theta}_{(1)}$. We denote by ω_i the normalised inverse variance weights for $\hat{\theta}_{(i)}$, i.e. $\omega_1 + \omega_2 + \cdots \omega_i = 1$. If the second study is conducted, then the combined effect $\hat{\theta}_{(2)}$ is

$$\hat{\theta}_{(2)} = \begin{cases} \omega_1\hat{\theta}_1 + \omega_2\hat{\theta}_2, & \text{with probability } p\left(\hat{\theta}_1, s_1^2, \theta_0\right), \\ \hat{\theta}_1, & \text{with probability } 1 - p\left(\hat{\theta}_1, s_1^2, \theta_0\right). \end{cases}$$

The main results on sequential design bias are given below in Equations (1–3), and their proofs are provided in Section A of the Supporting Information.

To derive the results on bias analytically, we assume that $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent and also that the weights are either non-random or at least independent of the estimated effects. This strong assumption, although common in meta-analysis, is fully satisfied for the weights based on inverse sample variances only when the effects are the sample means of continuous outcomes. It is also approximately true in the fixed-effect model when the studies in the meta-analysis are sufficiently large. To demonstrate that sequential decision bias arises in a quite general setting, in Section 2.2, we also provide simulation results for several decision-making models in random-effects meta-analysis. All our simulations are based on 10 000 values of $\hat{\theta}_1$, providing precision to the second decimal place for estimated biases.

Under the aforementioned assumptions, we prove that the expected value of the combined estimator is

$$E\left(\hat{\theta}_{(2)}\right) = \theta + (\omega_1 - 1)\mathrm{Cov}\left(p\left(\hat{\theta}_1, s_1^2, \theta_0\right), \hat{\theta}_1\right). \tag{1}$$

From Equation (1), we see, for example, that if the value of current estimate $\hat{\theta}_1$ is positively correlated with the probability of conducting an additional study, then (because $\omega_1 - 1$ is negative), the combined estimator $\hat{\theta}_{(2)}$ will be negatively biased.

A somewhat simpler version of our Equation (1) was obtained in equation (2.3) of Ellis and Stewart (2009) who considered equal weights and $p\left(\hat{\theta}_1, s_1^2, \theta_0\right) = p\left(\hat{\theta}_1 > cs_1^2\right)$, i.e. the second trial is carried out only if the result of the first trial is significant. This decision procedure is not intuitive and was labelled a 'toy story' by the authors.

Next, we study the conditional biases given the decision about the next trial. In practice, conditional biases are arguably more important than their unconditional counterparts. When a practitioner or a meta-analyst finds several trials in the literature, a particular decision-making scenario may have already taken place. Therefore, a standard meta-analytic estimate is, in fact, a conditional estimate.

Assume that $\mathrm{E}\left(p\left(\hat{\theta}_1, s_1^2, \theta_0\right)\right) \neq 0$ and let $Y = 1$ if the second trial is conducted and zero otherwise. Suppose Y and $\hat{\theta}_2$ are conditionally independent given $\left(\hat{\theta}_1, s_1^2\right)$. Then the conditional expectation given that the second trial is conducted is

$$E\left\{\hat{\theta}_{(2)} | Y = 1\right\} = \theta + \frac{\omega_1 \mathrm{Cov}\left(p\left(\hat{\theta}_1, s_1^2, \theta_0\right), \hat{\theta}_1\right)}{E\left(p\left(\hat{\theta}_1, s_1^2, \theta_0\right)\right)} + \omega_1\left\{E\left(\hat{\theta}_1\right) - \theta\right\}. \tag{2}$$

Thus, unless $\mathrm{Cov}\left(p\left(\hat{\theta}_1, s_1^2, \theta_0\right), \hat{\theta}_1\right) = 0$, both the unconditional and conditional expectations are biased. The last term in the preceding equation, although zero for an unbiased estimator $\hat{\theta}_1$, is retained intentionally, so that the equation can be generalised to the case of K sequential decisions and trials. Similarly, the conditional expectation given that the trial is not conducted is

$$E\left\{\hat{\theta}_{(2)} | Y = 0\right\} = \theta - \frac{\mathrm{Cov}\left(p\left(\hat{\theta}_1, s_1^2, \theta_0\right), \hat{\theta}_1\right)}{1 - E\left(p\left(\hat{\theta}_1, s_1^2, \theta_0\right)\right)}. \tag{3}$$

The consequence of Equations (2) and (3) is that when the probability of conducting an additional study is positively correlated with the value of the current estimate $\hat{\theta}_1$, the bias given that the study is conducted is positive, and negative given that it is not.

Remark 1

If K trials were run sequentially and the decision to run trial $i + 1$ was dependent on the cumulative results from the first i trials, $\hat{\theta}_{(i)} = \sum_{j=1}^{i} \omega_j \hat{\theta}_j$ for $i = 1, \cdots, K - 1$. Equation (2) can be applied directly to cumulative effect $\hat{\theta}_{(K-1)}$ and the effect in the K-th trial $\hat{\theta}_K$, to obtain a recurrent equation for the sequential decision bias

$$\mathrm{E}_K\left(\hat{\theta}_{(K)} - \theta\right) = \omega_{(K-1)}\mathrm{E}_{K-1}\left(\hat{\theta}_{(K-1)} - \theta\right) + \omega_{(K-1)}\mathrm{Cov}\left(p_{(K-1)}, \hat{\theta}_{(K-1)}\right)\left[\mathrm{E}\left(p_{(K-1)}\right)\right]^{-1}. \tag{4}$$

In Equation (4), $\mathrm{E}_i(\cdot)$ is the conditional expectation given i trials, and $\omega_{(i)} = \sum_1^i \omega_j / \sum_1^{i+1} \omega_j$ is the normalised weight for $\hat{\theta}_{(i)}$. Similarly, $p_{(i-1)} = p\left(\hat{\theta}_{(i-1)}, s_{(i-1)}^2, \theta_0\right)$ is the probability of running the ith trial.

2.2. Modelling the probability of the second trial

To examine the resulting biases numerically, a model for the probability of running the next trial $p\left(\hat{\theta}, s^2, \theta_0\right)$ given cumulative results $\left(\hat{\theta}, s^2\right)$ is required. We first examine three simple models: the power-law, the extreme value and the probit models, and then a more complex model depending on power calculations.

2.2.1. A power-law model for $p\left(\hat{\theta}, s^2, \theta_0\right)$

Suppose $\hat{\theta}_1 \sim N(\theta, \sigma^2)$ for $\theta > 0$, and, for some $t > 0$, $p\left(\hat{\theta}_1, s^2, \theta_0\right) = \left(\hat{\theta}_1 / \theta_0\right)^t$ for $0 < \hat{\theta}_1 < \theta_0$ and zero otherwise. That is, there is no need for further trials when the effect is at least θ_0, and 'promising' results increase the probability of the next trial. The variability of the estimator is not taken into account in this very simplistic model. The function $p(x, \theta_0)$ is a distribution function from the power-law family of distributions on $[0, \theta_0)$ and $t = 1$ corresponds to uniform distribution. In (15) in the Supporting Information, we see that the covariance between $\hat{\theta}_1$ and $p\left(\hat{\theta}_1, \theta_0\right)$ is negative if $\theta > \theta_0$, so we expect a negative bias for the conditional expectation in this case.

If $\theta_0 > \theta$, the bias may be positive. In Figure 1, we plot the heatmaps for the biases of the conditional and unconditional means as functions of $0 \le \theta$, $\theta_0 \le 1$ for $t = 3$, $\omega_1 = \omega_2 = 1/2$ and $\sigma^2 = 1/3$ (the value obtained as s_1^2/n_1 for the first trial in the example in Section 4). These heatmaps were computed by performing 10 000 simulations at each pair of values of θ, θ_0 in steps of 0.05. We note that the most biased estimator is the conditional estimator when a new trial is conducted. It can be seen that in this model, the unconditional mean is reasonably precise, but the step 2 conditional mean has a considerable positive bias when the actual effect is small in comparison with the target value of θ_0. The step 3 conditional mean appears to be even more biased for small values of the actual effect. Because the model assumes no need for a further trial when the effect value of θ_0 is reached, the model underestimates the mean for large values of *a*.

Figure 2 illustrates the biases arising in the random-effects model $\hat{\theta}_1 \sim N(\theta, \sigma^2 + \tau^2)$ for $0.3 \le \theta \le 0.7$ and the target value of $\theta_0 = 0.5$. There is much random variation in this figure due to the comparatively large variance of $\sigma^2 = 1/3$, but the biases are quite distinct. Plots (a) and (b) in this figure show that increases in τ^2 cause increased biases in the power-law model. Similar bias plots for the third trial ($i = 3$) following on from the second in the same fashion are given in the Supporting Information, Fig. S1, plots (a) and (b). Biases for $\sigma^2 = 0.04$ (corresponding to a sample size of 500) are given in Figs. S2 and S3 in the Supporting Information. Here, the impact of an increase in τ^2 is visible very clearly.

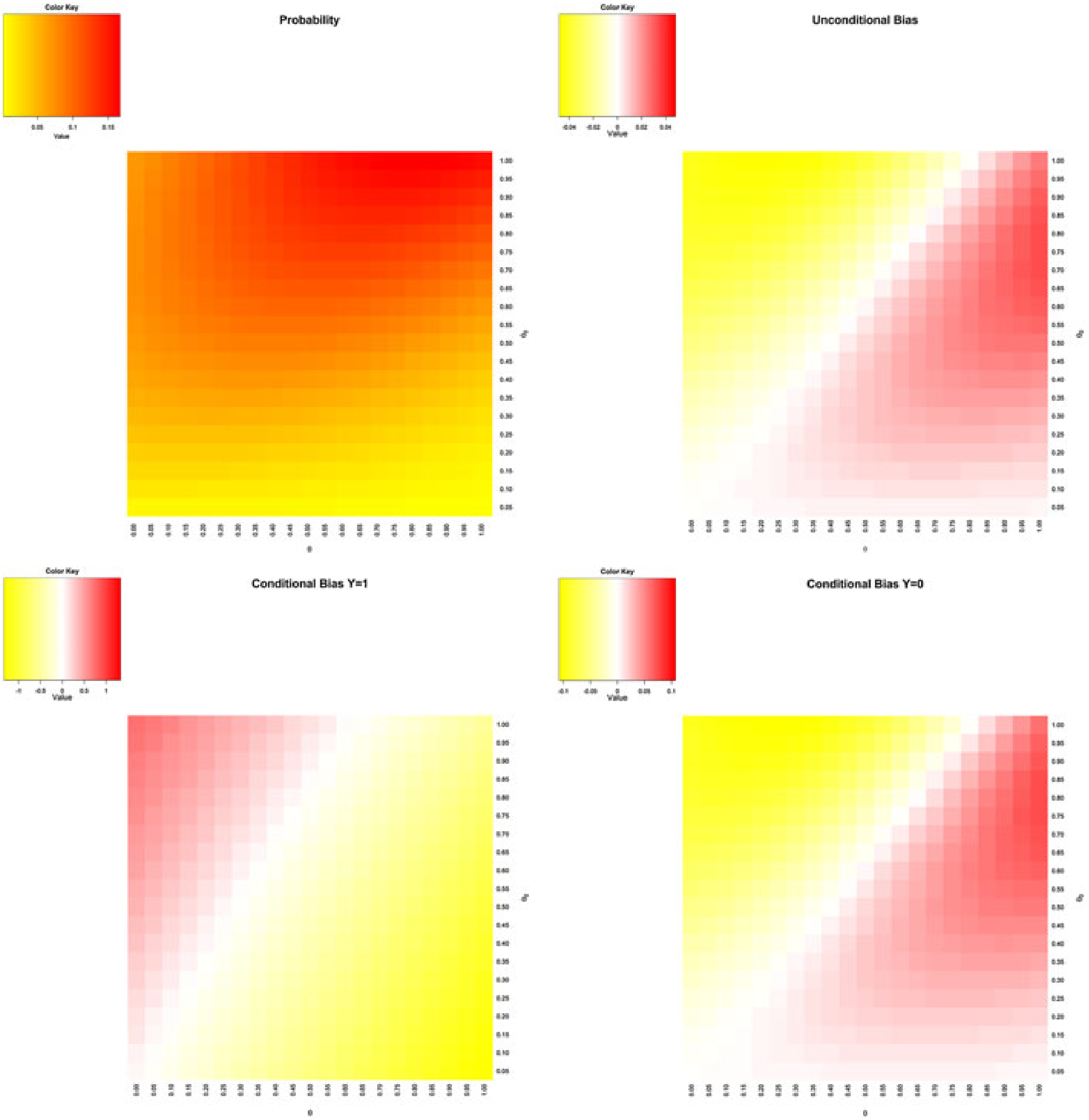


Figure 1. The probability of conducting a second trial and the bias in the unconditional and conditional means when the second trial is conducted (Y = 1) and not conducted (Y = 0), given by Equations (1), (2) and (3), respectively, for the power-law model of Section 2.2.1 with $t = 3$. The *x*-axis is the true value of θ whilst the *y*-axis is the target value θ. The parameter values are $\sigma^2 = 1/3$ and $\omega_1 = \omega_2 = 1/2$.

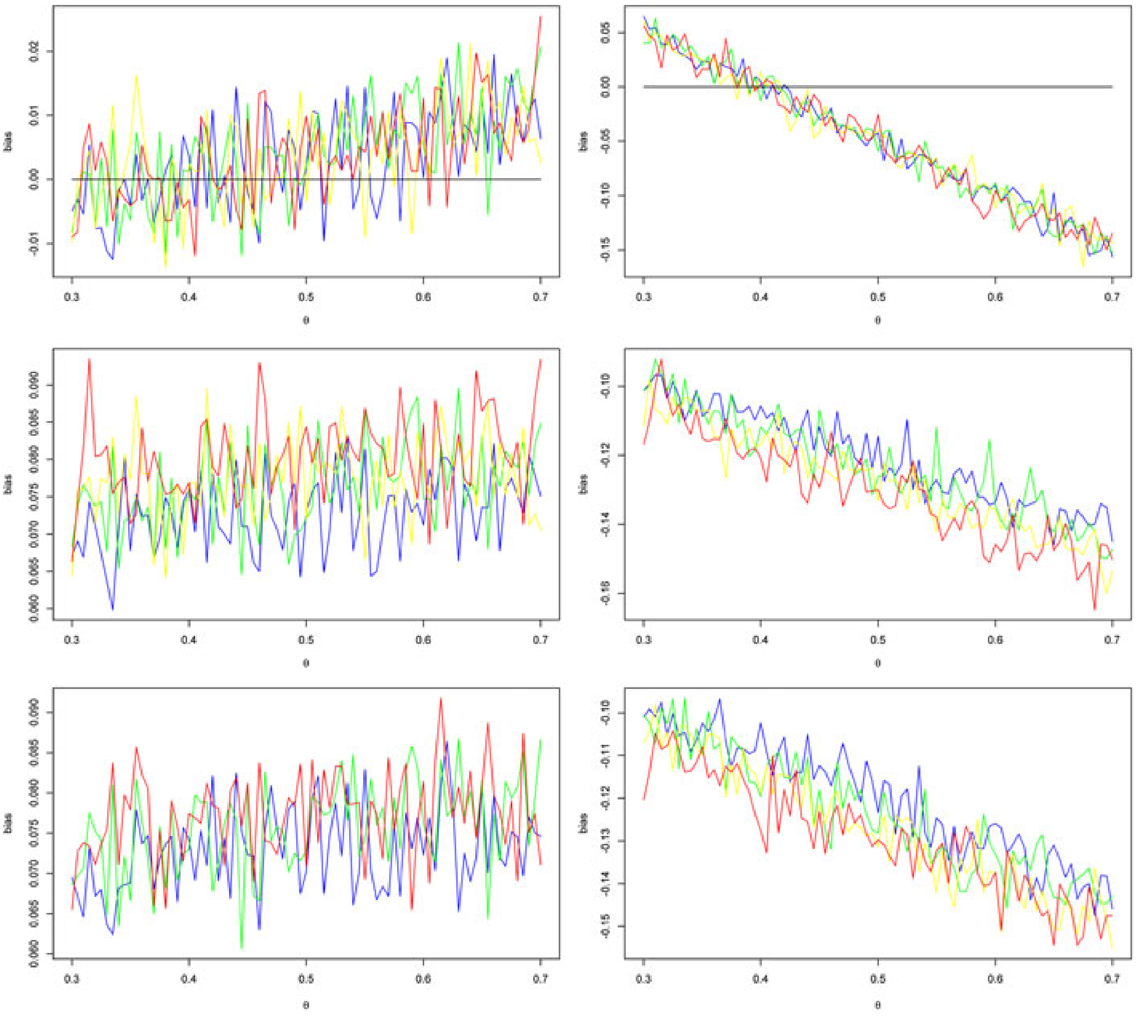


Figure 2. Biases of unconditional (left) and conditional $Y = 1$ (right) expected values of the cumulative effects $\hat{\theta}_{(2)}$ at the second study for τ^2 values of 0 (blue), 0.02 (green), 0.04 (yellow) and 0.06 (red). Rows 1 to 3 correspond to the biases in power-law with $t = 3$, extreme value ($r = 0.8$) and probit ($r = 0.8$) with $\alpha = 0$ and $\beta = 1$ models, respectively. Results from 10 000 simulations at each value of $\theta = 0.3(0.05)0.7$ for the target value of $\theta_0 = 0.5$, equal weights $\omega_1 = \omega_2$ and the variance $\sigma^2 = 1/3$ (corresponding to the within-study variance $s_1^2 = 19.94$ and sample size of $n_1 = 61$ in the example of Section 4).

2.2.2. Extreme-value and probit models. In this subsection, we briefly consider two alternative models for $p\left(\hat{\theta}, s^2, \theta_0\right)$. Both are based on a class of *t*-models for publication bias by Copas (2013). These models are of the form $a(\theta/\sigma)$ for an arbitrary function $a(\cdot)$. We centre these functions at the clinically significant effect θ_0 and truncate at $r\theta_0$ for some $0 < r < \theta_0$, so that the next trial is unlikely when the current estimated effect is much below (or much above) the clinically significant effect. The general form of these models is

$$p\left(\theta, \sigma^2, \theta_0, r\right) = [1 - G((\theta - \theta_0)/\sigma)]/[1 - G((r-1)\theta_0/\sigma)] \text{ for } \theta > r\theta_0,$$

and 0 otherwise, for a distribution function $G(\cdot)$. Consider first an extreme value distribution-based model for which the distribution function $G(\theta, \sigma^2, \theta_0) = \exp(-\exp((\theta_0 - \theta)/\sigma))$. The simulation-based bias plots for the second trial under this model with $r = 0.8$ are given in plots (c) and (d) of Figure 2. It can be seen that the unconditional expected values underestimate the mean, and conditional ($Y = 1$) expected values overestimate the mean. Biases noticeably increase with the increase in the heterogeneity parameter τ^2.

Following a probit model for publication bias by Copas (2013) we consider the probit model with $G(\theta, \sigma^2, \theta_0) = \Phi(\alpha + \beta(\theta - \theta_0)/\sigma)$. We provide simulation results for a simple version of this model with $\alpha = 0$ and $\beta = 1$ for $r = 0.8$ in the plots (e) and (f) of Figure 2. As with the extreme value model, the probit model also yields negative unconditional bias and positive conditional bias given that the next trial is conducted. Once more, the biases increase with an increase in the heterogeneity parameter τ^2.

The bias plots for the third trial ($i = 3$) are given in the Supporting Information, Fig. S1. It can be seen that the biases increase with each step i in decision-making.

As we have seen, different rules and different parameters could give quite different results, but these indicate that biases do occur when data-dependent rules are used to determine if the second trial should be conducted.

2.2.3. A power calculation model for $p\left(\hat{\theta}_1, s_1^2, \theta_0\right)$. In this section, we look at the situation in which the probability of conducting a new trial may depend on power calculations. For simplicity, suppose that two studies may be conducted and that we use normally distributed means to estimate an effect size θ that is the same for each trial. Typically, if the power calculations yield a small sample size for the second study, the increase in total power of the subsequent meta-analysis will be minor, and it may be decided that it is not worth proceeding with the study. Alternatively, the power calculations may yield a large sample size and it may not be possible to achieve the desired power with the available resources.

Let the first study result in an estimate $\hat{\theta}_1$ of θ. We wish to determine a sample size n_2 for the next study, so that the combined effect $\hat{\theta}_{(2)} = \left(w_1\hat{\theta}_1 + w_2\hat{\theta}_2\right)/(w_1 + w_2)$ will be significantly different from zero (two-sided) at the significance level α with $1-\beta$ power at the target effect size θ_0. Here, $w_i = n_i/\sigma_i^2$, $i=1,2$ are the unnormalised inverse variance weights, and σ_i^2 are the population variances within the studies. The level α should be chosen to account for multiple testing, but the details of such adjustments are beyond the scope of this paper. The variance of the combined effect is then $(w_1+w_2)^{-1}$. In the Supporting Information, we show that the required sample size is

$$n_2 = \left(\frac{c^2}{\theta_0^2} - w_1\right)\sigma_2^2, \tag{5}$$

where $c = z_{1-\alpha/2} + z_{1-\beta}$.

In the absence of clinical knowledge, the estimate of the treatment effect from the initial study is often taken to be the clinically relevant treatment effect for this sample size calculation – a practice we do not condone. In this scenario, the second sample size is calculated using the estimated mean $\hat{\theta}_1$ from the first trial as the effect size θ_0. The estimated variance s_1^2 may be used to estimate both σ_1^2 and σ_2^2 in the aforementioned formula. Then the sample size is taken to be

$$n_2 = \frac{c^2 s_1^2}{\hat{\theta}_1^2} - n_1. \tag{6}$$

If $\hat{\theta}_1$ is normally distributed and independent of the sample variance s_1^2, which has d_1 degrees of freedom ($d_1 = n_1 - 1$ for one sample, but we introduce this notation for more generality), then $d_1 s_1^2/\sigma_1^2 \sim \chi^2(d_1)$, and we may compute probabilities associated with the experiment. For example, suppose that it is decided to conduct a new study of size b if $a < n_2 \leq b$ for some a and b. Then assuming that $\hat{\theta}_1 \sim N\left(\theta, \sigma_1^2/n_1\right)$ is unbiased, the conditional (given $\hat{\theta}_1 = \theta_1$) and unconditional probabilities that a new trial is conducted are given by equations (18) and (19) in the Supporting Information Section A.

In Figure 3, we plot the estimated percentage unconditional bias as a function of θ from 10 000 simulated experiments for the same scenario with $n_1 = 60$, $\sigma_1^2 = 20$ and if a second experiment of size 80 is conducted when $30 \leq n_2 \leq 80$. The conditional bias is calculated over the simulations where a second trial was/was not conducted. The biases are very considerable, especially for small effects θ, where the conditional bias given that a new trial is conducted is above 100 %, and the negative unconditional bias is approximately 10 %.

In Fig. S4 in the Supporting Information, we plot the conditional and unconditional probabilities that a new trial is conducted for $n_1 = 60$, and $\sigma_1^2 = 20$, the parameters taken from the first trial in the Example discussed in Section 4. We assume $a = 30$ and $b = 80$ in these plots.

Remark 2

If the decision to run trial $i+1$ of size b is taken if $a < n_{i+1} \leq b$ for some $i \geq 1$, denote the cumulative combined effect from the first i trials $\hat{\theta}_{(i)} = \sum_{j=1}^{i} w_j\hat{\theta}_j / W_{(i)}$ for $W_{(i)} = \sum_{j=1}^{i} w_j$, and the cumulative sample size $n_{(i)} = \sum_{j=1}^{i} n_j$. Equation (5) changes to

$$n_{i+1} = \left(\frac{c^2}{\theta_0^2} - W_{(i)}\right)\sigma_{i+1}^2, \tag{7}$$

where σ_{i+1}^2 is the variance of the trial $(i+1)$. Equations (18) and (19) in the Supporting Information Section A can be adapted to provide the conditional and unconditional probabilities of the trial $(i+1)$ being conducted.

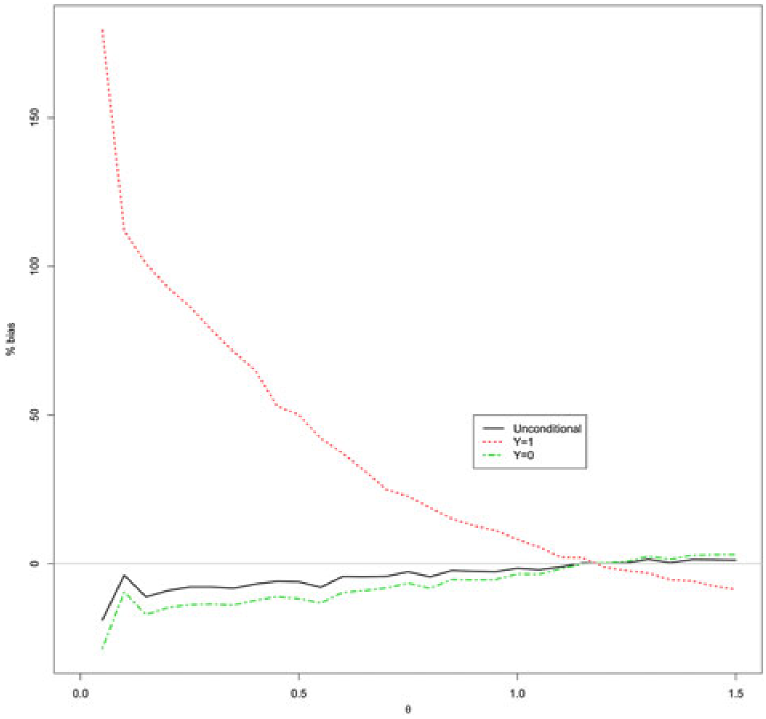


Figure 3. Estimated percent unconditional (solid line) and conditional (Y = 1 dashed line, Y = 0 dot-dashed line) bias from 10 000 simulations as a function of θ for the power calculation rule of Section 2.2.3, with $\sigma_1^2 = 20$, $n_1 = 60$, $a = 30$ and $b = 80$ and a second trial of size 80 is conducted if $a < n_2 < b$.

3. Sequential design bias

Although there is an implicit decision made prior to designing the second trial, we now emphasise the design of the second trial and refer to the resulting bias as design bias. We continue the approach and the notation of Section 3, but now rather than conduct a new experiment with b observations, we conduct it with n_2 observations. We investigate the bias introduced by sample size calculations based on estimated effects and their variances. For the ith trial, $i = 1, 2$, with n_i observations, the weight is $w_i = n_i/\sigma_i^2$ and the effect estimate $\hat{\theta}_i$ is taken to be a sample mean. The combined estimate over two trials is then $\hat{\theta}_{(2)} = \sum_{i=1}^2 w_i\hat{\theta}_i / \sum_{i=1}^2 w_i$. Note that $n_2 = 0$ yields $w_2 = 0$ and $\hat{\theta}_{(2)} = \hat{\theta}_1$. In what follows, we make the assumption that the estimates of the effect size and variances are independent, which will hold for samples from normal populations and approximately for other situations.

In practice, σ_2^2 is not known and we must guess a value to determine the sample size. Denote this by σ_g^2. A common option is to take $\sigma_g^2 = \hat{\sigma}_1^2$, which is explored later. Then, following from Equation (5), we take

$$n_2 = \left(\frac{c^2}{\theta_0^2} - w_1\right)\sigma_g^2 = \left(\frac{c^2}{\theta_0^2} - w_1\right)d^2\sigma_2^2,$$

where $d^2 = \sigma_g^2/\sigma_2^2$. Thus, n_2 is positive if $c^2 > w_1\theta_0^2$ or $n_1 < c^2\sigma_1^2/\theta_0^2$. Let $n_2 = \max(n_2, 0)$. Set $d = 0$ whenever $n_2 = 0$. Therefore, with $w_2 = n_2/\sigma_2^2 = \left(c^2/\theta_0^2 - w_1\right)d^2$, we have

$$\hat{\theta}_{(2)} = \frac{w_1\hat{\theta}_1 + w_2\hat{\theta}_2}{w_1 + w_2} = \hat{\theta}_2 + \frac{w_1\left(\hat{\theta}_1 - \hat{\theta}_2\right)}{w_1 + w_2}. \tag{8}$$

As long as the value of θ_0 used in the sample size calculation is a constant decided by clinical considerations, the expected value of the cumulative effect given by Equation (8) is equal to θ and is unbiased. But in the absence of this clinical knowledge, when designing the second study, it is tempting to use the value of $\hat{\theta}_1 + \delta$ for some constant δ for the sample size calculation. In clinical trials, this can form the basis of proceeding to a phase III trial. That is, we now have

$$n_2 = \left(\frac{c^2}{\left(\hat{\theta}_1 + \delta\right)^2} - w_1\right)\sigma_g^2 \tag{9}$$

and $\hat{w}_2 = \left(c^2/\left(\hat{\theta}_1 + \delta\right)^2 - w_1\right)d^2$. To examine this situation, for simplicity, suppose σ_1^2 and therefore w_1 too are known. Note that if $\hat{\theta}_1$ is large, then n_2 given by (9) can be negative, and in this case, a further experiment is not conducted.

In the Supporting Information Section A, we show that if we stop after the first experiment because the observed result had the desired power for the observed effect size, we can obtain a highly positively biased estimate of θ. For example, if $\theta = 0.2$, $\sigma_1^2 = 1$, $\alpha = 0.05$, $\beta = 0.2$, $\delta = 0.2$ and $n_1 = 15$, then $E(\theta_1 | n_2 \leq 0) = 0.647 >> 0.2 = \theta$. Fortunately, the bias diminishes with increase in n_1, so that for $n_1 = 50$, $E(\theta_1 | n_2 \leq 0) = 0.310$, and for $n_1 = 100$, $E(\theta_1 | n_2 \leq 0) = 0.222$. In the limit, the bias is zero. Similarly, given that the observed result has not reached the desired power and the second trial is conducted, the estimate of θ is negatively biased.

Now let us explore the unconditional bias of θ. Suppose that we guess the variance σ_2^2 exactly, i.e. $d = 1$. As a consequence of (20) in the Supporting Information Section A, we have

$$E\left(\hat{\theta}_{(2)}\right) \leq \theta + \left(\frac{\phi(h)\sigma_1}{\sqrt{n_1}} + \frac{2(\theta + \delta)}{c^2}\right), \tag{10}$$

where $h = \left\{\left(\sqrt{c^2/w_1} - \delta\right) - \theta\right\}/\left(\sigma_1/\sqrt{n_1}\right)$ and $\phi(\cdot)$ is the standard normal density, giving an upper bound on the unconditional bias of the estimate $\hat{\theta}_{(2)}$. That is, if $\delta > 0$ and $\alpha = 0.05$, $\beta = 0.2$ then $c^2 = 7.85$ and the bias is not greater than 25%. But what happens when $d \neq 1$?

In the general case, for an arbitrary d value,

$$\hat{\theta}_{(2)} = \hat{\theta}_2 + \frac{w_1\left(\hat{\theta}_1 - \hat{\theta}_2\right)\left(\hat{\theta}_1 + \delta\right)^2}{w_1\left(\hat{\theta}_1 + \delta\right)^2\left(1 - d^2\right) + d^2c^2}.$$

At $d = 0$, this is just $\hat{\theta}_1$ and is unbiased. However, this is of little practical use for at $d = 0$ we would not conduct the second experiment. Now, for an arbitrary d,

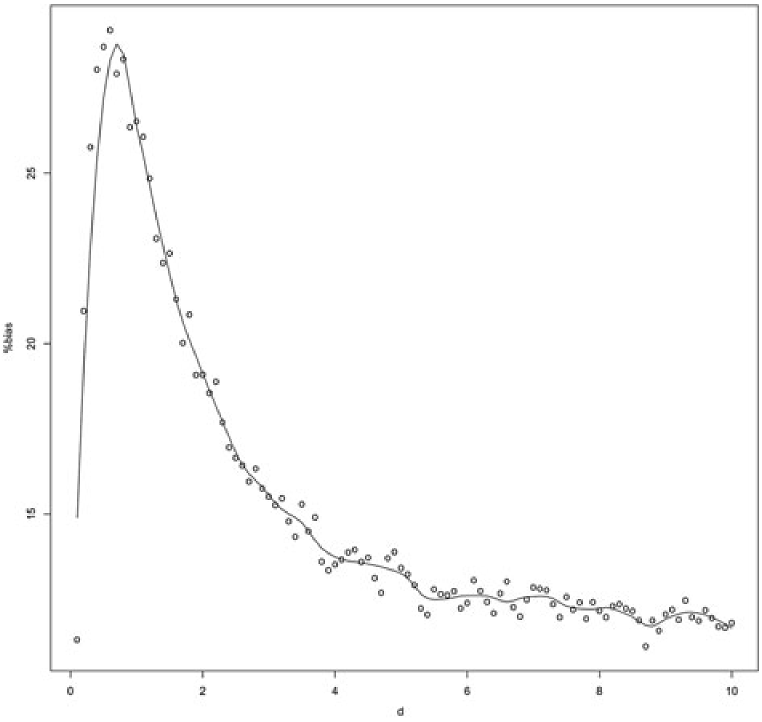


Figure 4. Plot of the percent bias against d from 10 000 simulations with $\theta = 0.2$, $\delta = 0.2$, $\alpha = 0.05$ and $\beta = 0.2$. The variances $\sigma_1^2 = 19.94, \sigma_2^2 = 24.96$ and the sample size $n_1 = 61$ were taken from the example of a meta-analysis discussed in Section 4. The locfit (Loader, 2012) package was used to smoothly estimate the mean bias.

$$E\left(\hat{\theta}_{(2)}\right) = \theta + E\left\{\frac{w_1\left(\hat{\theta}_1 - \theta\right)\left(\hat{\theta}_1 + \delta\right)^2}{w_1\left(\hat{\theta}_1 + \delta\right)^2\left(1 - d^2\right) + d^2c^2}\right\}, \tag{11}$$

which is analytically intractable.

We have seen that the bias at $d = 1$ can be quite large; therefore, we conducted a series of simulations to examine the bias for other values of d. The design of these simulations is described in the Supporting Information Section C.

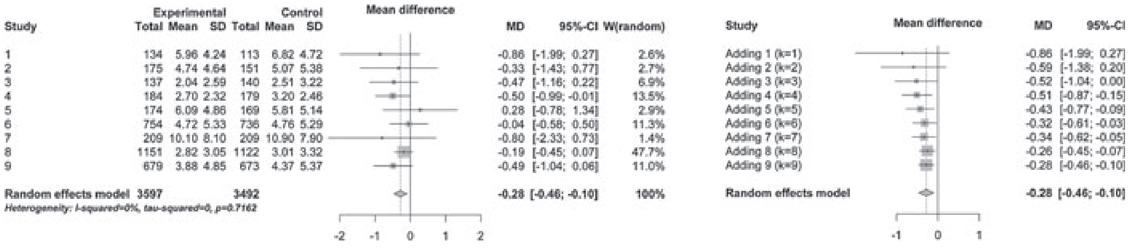
For our first simulations, we took $\theta = 0.2$, $\delta = 0.2$, $\alpha = 0.05$ and $\beta = 0.2$. The variances $\sigma_1^2 = 19.94$, $\sigma_2^2 = 24.96$ and the sample size $n_1 = 61$ were taken from the example of a meta-analysis discussed in Section 4 so that $w_1 = 3.06$. Then $c^2/\theta^2 - w_1 = 136.22 > 0$. We conducted 10 000 simulated initial experiments and took d from 0.1 to 10 in steps of 0.1. The means of the combined estimators for each value of d are plotted in Figure 4. As there is some variability due to the random sampling, we used the R (R Core Team, 2012) package locfit (Loader, 2012) to smoothly estimate the mean. It is clear from this plot that the bias can be substantial over a range of guesses for σ_2^2. With $\delta = 0$, the bias was around 15% less than that in Figure 4 but still of concern. If we took $\sigma_g^2 = \hat{\sigma}_1^2$, then the mean percentage bias over the simulations was 49.9% with standard deviation of 3.16 so the bias was uniformly high. As a check, if we used θ and σ_2^2 with $\delta = 0$ to compute n_2, then the mean bias was close to zero (0.129%). The bias was 0.042%, also close to zero if we used θ and $\hat{\sigma}_1^2$ again with $\delta = 0$ to compute n_2, which confirms that the bias arises from using the estimated value of $\hat{\theta}_1$ to decide to carry out the second experiment and compute the sample size.

4. Example

The meta-analysis conducted by Johnson (1993) comprised nine studies comparing sodium monofluorophosphate with sodium fluoride dentifrices in the prevention of caries development. The outcome of interest was the dental decay score, and a summary of the data, the forest plot, the cumulative meta-analysis plot and the results obtained with R package 'meta' (Schwarzer, 2010) are given in Figure 5. Heterogeneity was not detected, so the fixed-effect model was used. The first three studies in this meta-analysis failed to reach significance but showed positive effect. The combined effect after three trials, $\hat{\theta}_{(3)} = 0.5211$, is just significant ($p = 0.049$, confidence interval [0.003; 1.039]).

For purposes of an illustrative example, we will assume the following: that the meta-analyst could have been responsible first for deciding whether or not to conduct the second (and/or third) study given the results of the previous studies and second for deciding on the design of the additional study. We will examine the bias that would be introduced by these decisions.

All results formulated in the previous sections were based on just one sample size n_i and the sample variance s_i^2 per trial. Here, we first explain how to apply them to the standard clinical trial setting. Consider K comparative



Study	Experimental Total	Experimental Mean	Experimental SD	Control Total	Control Mean	Control SD	MD	95%-CI	W(random)
1	134	5.96	4.24	113	6.82	4.72	-0.86	[-1.99; 0.27]	2.6%
2	175	4.74	4.64	151	5.07	5.38	-0.33	[-1.43; 0.77]	2.7%
3	137	2.04	2.59	140	2.51	3.22	-0.47	[-1.16; 0.22]	6.9%
4	184	2.70	2.32	179	3.20	2.46	-0.50	[-0.99; -0.01]	13.5%
5	174	6.09	4.86	169	5.81	5.14	0.28	[-0.78; 1.34]	2.9%
6	754	4.72	5.33	736	4.76	5.29	-0.04	[-0.58; 0.50]	11.3%
7	209	10.10	8.10	209	10.90	7.90	-0.80	[-2.33; 0.73]	1.4%
8	1151	2.82	3.05	1122	3.01	3.32	-0.19	[-0.45; 0.07]	47.7%
9	679	3.88	4.85	673	4.37	5.37	-0.49	[-1.04; 0.06]	11.0%
Random effects model	3597			3492			-0.28	[-0.46; -0.10]	100%

Heterogeneity: I-squared=0%, tau-squared=0, p=0.7162

Study	MD	95%-CI
Adding 1 (k=1)	-0.86	[-1.99; 0.27]
Adding 2 (k=2)	-0.59	[-1.38; 0.20]
Adding 3 (k=3)	-0.52	[-1.04; 0.00]
Adding 4 (k=4)	-0.51	[-0.87; -0.15]
Adding 5 (k=5)	-0.43	[-0.77; -0.09]
Adding 6 (k=6)	-0.32	[-0.61; -0.03]
Adding 7 (k=7)	-0.34	[-0.62; -0.05]
Adding 8 (k=8)	-0.26	[-0.45; -0.07]
Adding 9 (k=9)	-0.28	[-0.46; -0.10]
Random effects model	-0.28	[-0.46; -0.10]

```
Number of studies combined: k=9

                         MD            95%-CI      z  p.value
Fixed effect model   0.2833  [0.1023; 0.4644] 3.0671   0.0022

Quantifying heterogeneity:
tau^2 < 0.0001; H = 1 [1; 1.38]; I^2 = 0% [0%; 47.7%]

Test of heterogeneity:
    Q d.f.  p.value
 5.38    8   0.7162
```

Figure 5. Forest plot (left), cumulative meta-analysis plot (right) and some additional results for Johnson(1993) meta-analysis.

studies with the treatment (T) and the control (C) arms of sample sizes N_{iT} and N_{iC}, $N_{iT}+N_{iC}=N_i$. The effect measure is the mean difference $\hat{\theta}_i=\overline{X}_{iT}-\overline{X}_{iC}$. Assume, for simplicity, equal variances σ_i^2 in both arms of each study. The variance of the effect is $\text{Var}\left(\hat{\theta}_i\right)=\sigma_i^2\left(N_{iC}^{-1}+N_{iT}^{-1}\right)$. This can be written as σ_i^2/n_i for the effective sample size n_i. This effective sample size is the geometric mean of the sample sizes of each arm, i.e. $n_i=\left(N_{iT}^{-1}+N_{iC}^{-1}\right)^{-1}$. For a balanced trial of size N_i, $n_i=N_i/4$. Now all the theory of Sections 2 and 3 can be applied using effects $\hat{\theta}_i$, effective sample sizes n_i and pooled sample variances s_i^2 with $d_i=N_i-2$ d.f. from each study.

For the first three trials, the effective sample sizes were $n=(61.30,\ 81.06,\ 69.24)$ and the pooled sample variances were $s^2=(19.94,\ 24.96,\ 8.56)$, respectively. Assuming standard choices of a 5% significance level and power of 80%, the constant c in Equations (6) and (9) is $c=z_{1-\alpha/2}+z_{1-\beta}=2.802$.

After the first trial, we consider performing the second trial with a goal of achieving a definitive meta-analysis. Suppose the second trial may run given that its sample size is between 200 and 2000 patients. This translates into the decision to conduct a new balanced trial if $a<n_2\leq b$ for $a=50$ and $b=500$. Using the estimated treatment effect $\hat{\theta}_1=0.86$, we calculate from (6) the value of $n_2=150.35$ required for this new trial. Assume that the true parameter values are $\theta=0.28$ (the combined effect from nine trials) and $\sigma^2=21.62$ (the pooled variance from nine trials). Under these assumptions, we estimate from equation (19) in the Supporting Information the unconditional probability of continuation after trial 1 to be $p=0.349$. If we do not wish to restrict sample sizes from below, take $a=0$, and the resulting estimated probability of continuation is $p=0.412$. To assess the resulting bias after two trials, we performed 10 000 simulated experiments with $n_1=61$ (effective sample size of the first trial), $\theta=0.28$ and $\sigma^2=21.62$ for $a=50$ and $b=500$. From the simulations, the estimated probability to continue after the first trial is 0.343 (in agreement with our theoretical estimate of 0.349), unconditional bias after two trials is $-19.92\,\%$, conditional bias given the decision to stop is $-34.40\,\%$ and conditional bias given the decision to continue is 8.13 %. Thus, if the meta-analyst had been responsible for deciding (based on the results of the first study) whether to conduct the second study, then a substantial sequential decision bias would have been introduced.

If the second trial was run first, the estimated effective sample size for the next trial would be $1719>b$, and the next trial would not be run. Now suppose that the first two trials were run independently from each other, but the decision is required about the third trial. The variances in these two trials are similar, so we proceed as suggested in Remark 2. The required sample size calculated from (7) is 376.02. The unconditional probability to continue is 0.25. Taking $a=0$ increases this probability to 0.30. As it happened, the third trial was run with an effective sample size of 69.24, resulting in marginal significance of the combined effect $\hat{\theta}_{(3)}$ of sodium monofluorophosphate.

So far, we considered the sequential decision scenario for this meta-analysis. To assess the sequential design bias in this realistic setting, suppose that after the first trial the investigators correctly assume that the effect is overestimated and use $\delta=-0.36$ to 'correct' the effect to 0.50 in the sample size calculation. We have performed 10 000 simulated experiments with $n_1=61$, $\theta=0.28$, $\sigma_1^2=19.94$, $\sigma_2^2=24.96$ and $\delta=-0.36$. The results are plotted in Fig. S5 in the Supporting Information Section C. For the values of $d\approx 1$, i.e. when the variance is guessed correctly, the bias is about 15 %. The bias is greatly reduced for the large values of d. This is to be expected as the large assumed variance would result in large sample size of the next trial.

It is clear that in this meta-analysis with moderate sample sizes, both types of biases are far from negligible.

5. Discussion

We have demonstrated theoretically and by simulations that both sequential decision bias and sequential design bias can arise in sequential and cumulative meta-analyses when the results of previous studies influence the design of a new study. This setting differs from the standard sequential meta-analysis setting in that a meta-analyst has an active role in the design of the subsequent trial aiming at a definitive meta-analysis. We have seen that both the conditional and unconditional biases can be non-negligible. Thus, caution needs to be exercised in conducting meta-analysis when prior knowledge has been used to design the trials being studied.

Our sample size calculations are based on the unconditional power of the Wald test for the combined effect. A recent paper by Roloff *et al.* (2013) advocated using the conditional power approach. We do not expect this to alleviate the bias. The design bias arising from the conditional power is discussed in the designed extension of a clinical trial setting by Denne (2000). In particular, that paper compared the biases of the estimated effects in conditional and unconditional setting and found that the differences were minor (see Figure 2 in Denne (2000)).

We considered both the fixed-effect and the random-effects models of meta-analysis and demonstrated analytically and by simulations that in both settings the problems due to sequential biases are apparent. According to our simulations, the sequential biases increase with increased heterogeneity.

Sequential meta-analysis results in inflation of type-I error due to multiple testing. This well-known issue was not discussed in any detail in this paper. A number of procedures aimed at adjustment of significance levels to safeguard the overall type-I error are available from a number of publications such as Pogue and Yusuf (1998), van der Tweel and Bollen (2010) and Higgins *et al.* (2011). Such adjustments will result in larger sample sizes of the new studies (formulae (6) and (9)) and may decrease sequential biases (formula (11)) through increases in critical values. Bias due to non-independence of studies as well as the timing of the meta-analysis itself was studied theoretically but without quantification by Ellis and Stewart (2009). Their findings are akin to the sequential decision bias discussed in Section 2.

Our models assumed that a new study is more likely if the existing evidence is in favour of a new treatment than if it is the other way around. However, we have not considered in detail how the interplay between the effect size and the uncertainty may affect the sequential biases. The value of information approach (Claxton & Sculpher, 2006; Claxton *et al*., 2002) is an alternative method to decide on the necessity of further research. This method is based on economic modelling comparing the costs involved in further research with benefits of reduction in uncertainty. This method is widely used in contemporary health policy decision-making (Claxton & Sculpher, 2006). It would be of great interest to investigate the existence of sequential decision bias resulting from this approach.

Note that in clinical trials, the favourable results of a phase II trial may be used to design the phase III trial, '*Estimates of treatment effects and variability from earlier trials are traditionally used in the design of trials at the next stage*' (Kirby *et al*., 2012). This setting is different from meta-analysis in that results are not combined. Moreover, the decision to conduct the phase III trial depends on a significant result in the phase II trial, whereas with meta-analysis guiding research, the sequence of trials may be terminated once significance is attained. However, the problem of resulting biases is already recognised in drug development, (Wang *et al*., 2006) and methods of adjustment are sought (Kirby *et al*., 2012). Perhaps a closer analogy for the sequential decision bias is with group-sequential clinical trials, where a significant result at an interim stage would stop the trial, but otherwise, the results of sequential interim stages are accumulated and combined. In this setting, the existence of sequential bias is widely recognised and the means of adjustment for this bias have been developed (Whitehead, 1986). This adjustment is possible because of the explicit decision rules in these trials. Design bias is similar to the bias induced by mid-trial sample size re-estimation in adaptive trials (Li *et al*., 2002; Wang *et al*., 2010). However, methods of sequential bias adjustment in meta-analytic setting are more difficult to develop than in sequential and adaptive clinical trials. The bias depends not only on the unknown true value or the precision of the effect θ but also on the strategy for making the decision to continue or stop and of choosing the sample size of the next study. If such a strategy is made explicit, by, say, the Research Councils, development of an appropriate bias adjustment should be possible. Development of such a strategy appears to be an important and complicated problem deserving concerted efforts of statisticians and decision-makers.

As was pointed out by the Associate Editor, there are clear parallels between the models for the sequential decision bias and the models for publication bias, although the decisions being made are aimed at the next trial or at the current trial, respectively. In fact, both the model by Ellis and Stewart (2009) and our extreme-value and probit models are the publication bias models from the *t*-family introduced by Copas (2013). Similar to publication bias models, the choice of the model may greatly affect the amount of bias, but we have demonstrated that non-negligible biases may arise in a variety of models, and we recommend the use of multiple models for sensitivity analyses to uncover the consequences of sequential biases. To enable this use, all our R programmes are provided in the Supplementary Information materials.

We believe that this is the first time that this important issue is raised in the context of the sequential decision-making associated with the managed accumulation of evidence. Existence of sequential biases raises a number of important research questions. What is the best way to decide on the usefulness of a new trial? How to design this trial so that the resulting combined estimate is the least biased? How to adjust the combined effect to minimise this bias? All these questions need to be addressed if we are to aim at evidence-based development of science.

Acknowledgements

The work by the first author was supported by the Economic and Social Research Council (grant number ES/L011859/1).

The authors thank Prof M. B. Dollinger, two anonymous referees and an associate editor for helpful suggestions on the presentation of the contents of this article.

References

Claxton K, Sculpher M. 2006. Using value of information analysis to prioritise health research: some lessons from recent UK experience. *PharmacoEconomics* **24**: 1055–1068.

Claxton K, Sculpher M, Drummond M. 2002. A rational framework for decision making by the National Institute for Clinical Excellence (NICE). *The Lancet* **360**: 711–717.

Cooper N, Jones D, Sutton A. 2005. The use of systematic reviews when designing studies. *Clinical Trials* **2**: 260–264.

Copas JB. 2013. A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Journal of the Royal Statistical Society: Series C: Applied Statistics* **62**: 47–67.

Denne J. 2000. Estimation following extension of a study on the basis of conditional power. *Journal of Biopharmaceutical Statistics* **10**: 131–144.

Ellis S, Stewart J. 2009. Temporal dependence and bias in meta-analysis. *Communications in Statistics - Theory and Methods* **38**: 2453–2462.

Emerson S, Fleming T. 1990. Parameter estimation following group sequential hypothesis testing. *Biometrika* **77**: 875–892.

Glasziou P, Djulbegovic B, Burls A. 2006. Are systematic reviews more cost-effective than randomised trials? *Lancet* **367**: 2057–2058.

Goudie A, Sutton A, Jones D, Donald A. 2010. Empirical assessment suggests that existing evidence could be used more fully in designing randomized controlled trials. *Journal of Clinical Epidemiology* **63**: 983–991.

Higgins J, Whitehead A, Simmonds M. 2011. Sequential methods for random effects meta-analysis. *Statistics in Medicine* **30**: 903–921.

Johnson M. 1993. Comparative efficacy of NaF and SMFP dentifrices in caries prevention: a meta-analytic overview. *Caries Research* **27**: 328–336.

Kirby S, Burke J, Chuang-Stein C, Sin C. 2012. Discounting phase 2 results when planning phase 3 clinical trials. *Pharmaceut. Statist.* **11**: 373–385.

Li G, Shih W, Xie T. 2002. A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* **3**: 277–287.

Loader C. 2012. locfit: local regression, likelihood and density estimation. *R package version* **1**: 5–8.

Mulrow C. 1994. Systematic reviews: rationale for systematic reviews. *BMJ* **309**: 597–599.

Pogue J, Yusuf S. 1998. Overcoming the limitations of current meta-analysis of randomised controlled trials. *The Lancet* **351**: 47–52.

R Core Team. 2012. R: a language and environment for statistical computing. ISBN: 3-900051-07-0.

Roloff V, Higgins J, Sutton A. 2013. Planning future studies based on the conditional power of a meta-analysis. *Statistics in Medicine* **32**: 11–24.

Schwarzer G. 2010. meta, v.1.5-0. *CRAN* R package. Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot.

van der Tweel I, Bollen C. 2010. Sequential meta-analysis: an efficient decision-making tool. *Clinical Trials* **7**: 136–146.

Wang S, Hung H, O'Neill R. 2006. Adapting the sample size planning of a phase III trial based on phase II data. *Pharmaceut. Statist.* **5**: 85–97.

Wang Y, Li G, Shih W. 2010. Estimation and confidence intervals for two-stage sample size flexible design with LSW likelihood approach. *Statistics in Biosciences* **2**: 180–190.

Whitehead J. 1986. On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73**: 573–581.

Whitehead A. 1997. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in Medicine* **16**: 2901–2913.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.