Pharmaceutical Statistics

A note regarding meta-analysis of sequential trials with stopping for efficacy

Journal:	*Pharmaceutical Statistics*
Manuscript ID:	PST-13-0110.R2
Wiley - Manuscript type:	Main Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Senn, Stephen; CRP-Sante, Competence Center for Methodology and Statistics
Key Words:	Meta-analysis, Sequential analysis, Fixed effects, Weighting, Information, Consistency
Abstract:	It is shown that fixed-effect meta-analyses of naïve treatment estimates from sequentially run trials with the possibility of stopping for efficacy based on a single interim look are unbiased (or at the very least consistent, depending on point of view) provided that the trials are weighted by information provided. A simple proof of this is given. An argument is given suggesting that this also applies in the case of multiple looks. The implications for this are discussed.

A note regarding meta-analysis of sequential trials with stopping for efficacy

Stephen Senn

Abstract

It is shown that fixed-effect meta-analyses of naïve treatment estimates from sequentially run trials with the possibility of stopping for efficacy based on a single interim look are unbiased (or at the very least consistent, depending on point of view) provided that the trials are weighted by information provided. A simple proof of this is given. An argument is given suggesting that this also applies in the case of multiple looks. The implications for this are discussed.

1. Introduction

There seems to be some disagreement amongst researchers concerning the effect on the validity of a meta-analysis of sequential stopping rules in individual trials. Ioannidis has suggested that it could be one amongst many explanations as to why research findings are exaggerated[1]. Hughes et al have suggested that such stopping leads to biased results[2]. However Todd showed via simulation that any bias in a meta-analysis arising from combining trials that had been conducted sequentially, was likely to be small[3] and Schou and Marschner have recently completed an investigation with similar conclusions[4]. This is, of course, related to a much longer-running debate about the effect of stopping rules on inference. See, for example, Royall[5] section 5.3. This brief note provides a simple investigation of the circumstances under which unbiasedness is likely to be true (when there is a single interim look), shows that it is sufficient that the trials are weighted according to information provided and provides a simple proof of this claim. The context of early stopping for efficacy of individual trials is implicitly assumed. A referee has pointed out to me that other forms of stopping (for example futility) can be envisaged. I believe that the argument carries over to such cases but, in order to keep the argument simple, they are not considered. Having dealt with the case of a single look, an argument is then given as to why this finding will also apply for multiple looks.

Note that there is another sense in which sequential analysis is relevant to meta-analysis. This is when the intention is to stop running trials once a meta-analysis is significant. In other words it is the meta-analysis that is sequential rather than the individual trials. This problem has been studied by many authors[6-9] . However, it is a different issue and is not discussed in this note.

The structure of this note is as follows. In the next section I give a simple analogy to show that sequential decision-making can have an important influence on the expectation of some statistics but not on others. This is to encourage the reader to believe it is at least possible that a procedure that affects significance of tests does not affect expectation of estimates.

Then, in section 3 a simple model is introduced, the implications of which are considered in section 4 and illustrated numerically with section 5, which also provides the results of a simulation. Section 6 considers a possible objection to do with the meta-analysis of a single trial. Section 7 considers the extension to multiple looks and section 8 offers a discussion.

2. An analogy

Consider an imaginary world in which all couples decide to stop having children once they have a first son or four children, whichever is first. Assume that all couples can and will have four children if necessary. Assume that the probability of any male child is θ. The probabilities of given completed families are as follows

Family	Probabilty	Expected Males	Expected Children
M	θ	θ	θ
FM	$(1-\theta)\theta$	$(1-\theta)\theta$	$2(1-\theta)\theta$
FFM	$(1-\theta)^2\theta$	$(1-\theta)^2\theta$	$3(1-\theta)^2\theta$
$FFFM$	$(1-\theta)^3\theta$	$(1-\theta)^3\theta$	$4(1-\theta)^3\theta$
$FFFF$	$(1-\theta)^4$	0	$4(1-\theta)^4$
Overall	1	$1-(1-\theta)^4$	$\frac{1-(1-\theta)^4}{\theta}$

(1)

The distribution of different types of family is very strange reflecting the stopping rule for family completion. For example a completed family with a single daughter is impossible whereas for values of $\theta > 0.276$ a one son family is the most probable sort. Despite this, however, the sex ratio is

$$\frac{1-(1-\theta)^4}{\left[1-(1-\theta)^4\right]/\theta} = \theta$$

and so is completely unaffected by the sequential strategy, a point that is closely related to the properties of martingales and the impossibility of gambling strategies to affect expected winnings.

As will be shown below, we have with sequential trials an analogous situation, whereby although the expected values of certain statistics are indeed affected by the stopping rule, others such as for example the mean treatment effect (weighted by information) are not.

3. A model

To understand the general approach of the argument of this and the following section it will be helpful if the reader bears the following in mind. In general, one is worried that *stopped* trials will tend to *overestimate* the treatment effect. However, it thus follows that *completed* trials will tend to *underestimate* the treatment effect. Thus, a suitable linear combination of results from stopped and completed trials ought to be unbiased. The trick is to find the suitable linear combination.

For simplicity, consider a set of trials run according to a common protocol. A referee has rightly requested that I clarify the importance of this assumption. In fact it is only needed to make the argument simple. If different protocols are involved we just carry out a separate meta-analysis for each set of similar protocols as a preliminary step to combining these meta-analyses. The argument that follows is an argument in *expectation* and applies for any such set *even if there is only one trial in the set*. (This point is discussed further in section 6 below.) Once a separate valid analysis for each type of protocol has been performed these separate analyses can be combined.

Consider such a set of trials with a common protocol and a single look. If run to completion such trials will provide information which we may arbitrarily set equal to 1. An interim look will be carried out when the information fraction is $f, 0 < f < 1$. We let $X = 0,1$ be a random variable denoting whether a trial will stop ($X = 1$) or complete ($X = 0$). We let $P = E[X]$ be the probability of stopping. We suppose that all trials would provide an unbiased estimate of a common treatment effect τ if there were no optional stopping. We now consider three statistics: $\hat{\tau}_{s1}, \hat{\tau}_{c1}, \hat{\tau}_{c2}$. The first of these is the estimate of τ provided by a stopped trial, the second is the estimate provided by the first portion, corresponding to an information fraction f of a completed trial and the third is the portion corresponding to the remaining fraction, $1-f$ of a completed trial. We suppose that these three statistics have the following expectations

$$\begin{aligned} E\left[\hat{\tau}_{s1} \middle| X = 1\right] &= \tau + \delta_s \\ E\left[\hat{\tau}_{c1} \middle| X = 0\right] &= \tau + \delta_c \\ E\left[\hat{\tau}_{c2}\right] &= \tau. \end{aligned} \tag{2}$$

Here, only the third of the statistics in (2) is an unbiased estimate of τ since it is based on data collected subsequent to any decision used to stop a trial and these data are therefore not used for making that decision. The first two are conditional expectations since $\hat{\tau}_{s1}$ is only observed if $X = 1$ and $\hat{\tau}_{c1}$ is only observed if $X = 0$.

Now suppose that we define a further composite random variable

$$\hat{\tau}_1 = X\hat{\tau}_{s1} + (1 - X)\hat{\tau}_{c1} \tag{3}$$

This is simply a representation of the result we would expect to see after an information fraction of f without knowledge of whether the trial stopped or continued. Hence we have that the expectation of the left hand side of (3) is $E\left[\hat{\tau}_1\right] = \tau$. However, working with the right hand side of (3) we have

$$\begin{aligned}\tau &= E\left[X\hat{\tau}_{s1} + (1-X)\hat{\tau}_{c1}\right] \\ \tau &= E\left[X\hat{\tau}_{s1}\right] + E\left[(1-X)\hat{\tau}_{c1}\right] \\ \tau &= E[X]E\left[\hat{\tau}_{s1}\middle|X=1\right] + E[1-X]E\left[\hat{\tau}_{c1}\middle|X=0\right] \\ \tau &= P(\tau+\delta_s) + (1-P)(\tau+\delta_c) \\ \tau &= \tau + P\delta_s + (1-P)\delta_c.\end{aligned} \tag{4}$$

Therefore, equating left hand and right hand sides of (4) it follows that we have

$$P\delta_s + (1-P)\delta_c = 0. \tag{5}$$

4. Implications of the model

Suppose that we could separately identify results from the first portion of completed trials. If we simply weighted *such portions* the same as stopped trials then in any set of trials run under a sequential protocol we would expect to have a proportion P of stopped trials and $(1-P)$ of the corresponding fractions f of completed trials. It thus follows from (5) that the expected bias from the average of these trials is $P\delta_s + (1-P)\delta_c = 0$. Note that provided that the statistics $\hat{\tau}_{s1}$, which will arise with probability P and $\hat{\tau}_{c1}$, which will arise with probability $(1-P)$, are weighted equally, it does not matter what weight we give the statistics, $\hat{\tau}_{c2}$, corresponding to the remaining portion of the completed trial; the resulting combined estimator will be unbiased. Note, however, that it is *necessary* that $\hat{\tau}_{s1}$ and $\hat{\tau}_{c1}$ are given the same weight.

However, in practice, conditions of efficiency dictate that we ought to weight $\hat{\tau}_{c1}, \hat{\tau}_{c2}$ according to the information fractions $f, (1-f)$ to produce a combined statistic

$$\hat{\tau}_c = f\hat{\tau}_{c1} + (1-f)\hat{\tau}_{c2}. \tag{6}$$

In any case, in most meta-analyses portions $\hat{\tau}_{c1}, \hat{\tau}_{c2}$ will not be separately identifiable. Now suppose we consider the issue of correctly weighting $\hat{\tau}_{s1}$ and $\hat{\tau}_c$ so as to produce an unbiased estimate. If we weight $\hat{\tau}_{s1}$ proportionately to f and $\hat{\tau}_c$ proportionately to $f + (1-f) = 1$, that proportion of $\hat{\tau}_c$ that

is equal to $\hat{\tau}_{c1}$ will be given exactly the same weight as $\hat{\tau}_{s1}$, which is exactly the condition required for unbiasedness of the meta analysis. In other words we require to weight all stopped trials proportionate to the information fraction at stopping, that is to say by $f/(1+f)$ and all completed trials by $1/(1+f)$. More generally, if we consider a mixture of sequential trials, each run according to a different stopping rule we can see that all that is necessary is that each random variable representing the result from any type of trial is weighted proportionally to the information provided. However, this is exactly what happens in a fixed effects meta-analysis.

5. A numerical illustration

Figures 1 and 2 give the expected treatment estimate that will occur from meta-analyses performed on sequential trials. In each case it is assumed that the standard deviation of a Normally distributed estimate, X of the treatment effect for a trial that runs to conclusion is 1 and for each case the information fraction at which stopping can occur is varied in 99 steps from 0.01 to 0.99. A naïve (that is to say unadjusted) boundary, a, for a 5% stopping rule is used. For simplicity the standard deviation is assumed known. The expected means for stopped and continued trials are calculated using formulae for the truncated Normal distribution [10]. For the stopped trial the expected mean is

$$E\left(X \middle| X \geq a\right) = \tau + \frac{\sigma\phi\left(\frac{a-\tau}{\sigma}\right)}{1-\Phi\left(\frac{a-\tau}{\sigma}\right)} \tag{7}$$

and for the continued trial it is

$$\left(X \middle| X < a\right) = \tau - \frac{\sigma\phi\left(\frac{a-\tau}{\sigma}\right)}{\Phi\left(\frac{a-\tau}{\sigma}\right)}, \tag{8}$$

where $\phi(.)$ is the probability density function of a standard Normal distribution and $\Phi(.)$ is the cumulative density function. Note that both a and σ depend on the information fraction and note also that the stopped trial, for which (7) applies, occurs with probability $1-\Phi\left((a-\tau)/\sigma\right)$ and the continued trial, for which (8) applies, occurs with probability $\Phi\left((a-\tau)/\sigma\right)$, from which the unbiasedness property claimed in sections 3 and 4 follows trivially.

For each figure, two approaches to weighting are used. The first weights trials by information (as in a conventional meta-analysis) and the second weights trials equally. In addition to the theoretical figures, the results of simulating 4000 trials at each stopping fraction are shown.

<Figure 1 about here>

The first figure shows the situation for a treatment effect of 0, that is to say under the null hypothesis, whereas the second shows the situation for a true treatment effect of 1. In both cases the figures show that a weighted meta-analysis produces an unbiased estimate of the treatment effect for any information fraction. On the other hand, both figures also show that an un-weighted meta-analysis produces a biased estimate with the lower the information fraction at which the trials can be stopped the higher the bias. For an information fraction of 0.5, the expected value for the un-weighted estimator is 0.073 for the null case (Figure 1) and 1.18 when the true value is 1 (Figure 2).

<Figure 2 about here>

6. The meta-analysis of a single trial

A referee has raised a disturbing paradox. From one point of view, a simple analysis of a single trial is a meta-analysis. Yet for the argument of sections 3 and 4 to carry through it is necessary to be able to weight portions of a trial according to the information they carry. In particular a continued trial must be given more weight than a stopped trial. How can such weighting take place if one only has one trial? If weighting cannot take place, then in *expectation* one simply has an average, according to the probability with which they arise, of the mean from a stopped and a continued trial. In that case we have an un-weighted average of the stopped and continued trials and that, as we have already shown, is *not* unbiased.

I see three defences to this argument. First, one can say that one must think of a trial in terms of the information it provides and accept that not all trials are equal in this respect. Consider the example of a sequential analysis with a single look. We regard a continued trial as consisting of two trials: the first part in which stopping was possible and the second in which it is not. The second does not worry us. We know that information from the second part is unbiased. The question is, 'is the information from the first part unbiased?'. Here, the answer is , 'yes'. The situation is illustrated graphically in Figure 3, which shows the probability densities for the first part and the second part of the trial and how they are spliced together in probability. For this first part of the trial, the part that is potentially worrying because elective stopping may take place, one can then say that the estimate is unbiased. Thus unbiasedness can be claimed to apply to both parts in the following way: it applies to that part of the trial in which information is obtained under elective stopping and it applies (trivially) to that part where no such stopping takes place. Whether it applies to the trial *as a whole* is then a moot point. I regard the ambiguity as being a weakness of the classical notion of unbiasedness but accept that given the conventional definition of unbiasedness there may be a problem.

<Figure 3 about here>

The second defence is to say that if we let the number of trials go to infinity the problem disappears (because weighting becomes possible) so that the estimate is at least consistent (in increasing number of trials) if not unbiased.

The third defence is to make appeal to a concept of Bayesian unbiasedness. Provided the information can be combined with some prior information (however vague), then weighting *does indeed* take place, since trial results are combined with prior information. A stopped trial will have less influence on the posterior distribution than a continued one because the continued one has more information. Bayesian estimates are thus weighted estimates and thus Bayesian analysis permits the weighted combination of trials (in expectation by proxy) to take place.

7. Extension to the case with multiple looks

John Whitehead has pointed out to me that it does not necessarily follow from a proof that a meta-analysis of sequentially runs trials with a single look will produce unbiased estimates, that this will also be the case for trials with multiple looks. In this section an argument is sketched as to why a conventional fixed effects meta-analysis of such trials is unbiased.

Consider a trial run to a given protocol and which may be stopped at $look_j \;\; j = 1 \cdots k$ each look following a *stage* indexed identically. (Thus we can speak if we wish of $look_j$ and also of $stage_j$.) These looks do not have to be at equal intervals, whether in time or information. We now note that any estimate of the effect from the trial as a whole corresponds to a combination of the information from each section. It is thus sufficient for the estimate of the effect from the trial as a whole to be unbiased if a) the estimate from each stage is unbiased and b) the weights applied in combining such estimates across stages are independent of the values of the stages.

Now consider the estimate from $stage_j$. Such an estimate is unbiased. To see this it is simply necessary to note that conditional on the true parameter value everything that happened before the stage started is irrelevant. The increment of information from that stage is independent of all previous stages and also, therefore, of all subsequent ones. (This follows because, applying the argument to $stage_{j+1}$ it is independent of $stage_j$ and hence *vice versa*.)Here it is necessary to avoid the trap of thinking that one is looking at a trial that *will* stop at $look_j$. It *may* stop at $look_j$ it may not. All that is necessary is that the average over all such possible histories for the given stage is equal to the true parameter value. This is simply guaranteed by weighting all such estimates equally.

The question that then arises is, 'does a fixed -effects meta-analysis weight such estimates appropriately?'. Now consider any trial run under this protocol stopped at any stage whatsoever. We can, when faced with such a trial, re-arrange all the data to that we have two estimates: one, $\hat{\tau}_j$ (say) based on the stage we are considering and another, $\tau_{\neg j}$ based on all other stages combined (of which there might be none). It is not actually necessary to the argument but makes it easier to see if we now consider the largest possible information unit such that the information accruing during any possible stage of the trial can be expressed as an integer multiple of this. Suppose that the multiple for $stage_j$ is m_j. We now note that m_j is identical for all trials that give information in $stage_j$. Thus the information accruing to $\hat{\tau}_j$ is identical irrespective of what else happened in the trial

It is not the case that that the information is identical for τ_{-j} irrespective of what happened in the trial but this does not matter. We simply allow that the total information for the trial is the sum of the information for all the stages that were actually completed. (If one wishes we can regard each trial as being a meta-analysis of the stages weighted by information fractions.) If we now weight each trial in terms of the total information it produces, the contribution for $stage_j$ to the total is always identical provided that the stage was reached. This means that the contribution to the overall estimate from this stage is unbiased. But since $stage_j$ is any stage, the contribution from each and every stage is also unbiased and hence the estimate for the meta-analysis as a whole is unbiased.

8. Discussion

The result presented here may seem paradoxical. In particular, one could raise the following example. Suppose that a meta-analyst carries out a meta-analysis of a number of trials run to a sequential model and they all just happened to have stopped early. The result proved here says that on average over all such meta-analyses the estimated treatment effect equals the true unknown effect. But is this long-run average justification reassuring *given* that all of the trials have stopped early? Would the meta-analyst not be faced with a recognisable subset, one to which the long-run average does not apply?

I think not, but one has to very careful here. For instance one could certainly show that given a known value of τ an early stopped trial will not provide an unbiased estimate of τ. However, in estimation one does not have knowledge of τ and so the means by which a subset could be recognised may not be present. Of course a Bayesian has a prior belief in τ and the posterior mean (as discussed in section 6 above) will be a weighted combination of the data mean and the prior mean with greater shrinkage occurring in stopped trials because they are smaller[11]. There is, indeed, a sense in which frequentists and Bayesians ought to believe that results from sequential trials are less reliable than results from trials run to a standard fixed size. This is because the whole point about sequential trials is that they should deliver less information: they are planned with the hope that a reasonable decision may be reached earlier. I think that this particular point has sometimes been overlooked in the discussion of the currently fashionable topic of flexible designs. Such designs are meant to deliver less information and merely ensuring that the type I error rate is protected in doing this does not mean that the ability to satisfy other purposes of clinical trials is not harmed[12]. For examples of some paradoxes of inference that can arise is running such trials see Burman and Sonneson[13].

Finally, there are some intriguing issues raised regarding variance estimation. The first is that it is well-known, even for a meta-analysis of trials with fixed size, that confidence limit coverage is not maintained if (as is always the case) the variance of the treatment effect has to be estimated[14-16]. Is the situation worse for meta-analysis of sequential trials? Could it be the case that sequential stopping could increase the heterogeneity of results as was discussed in Hughes et al[2]? Would such heterogeneity make it more likely that a random effects approach be used? Would such an approach be biased whereas the fixed effects approach was not?

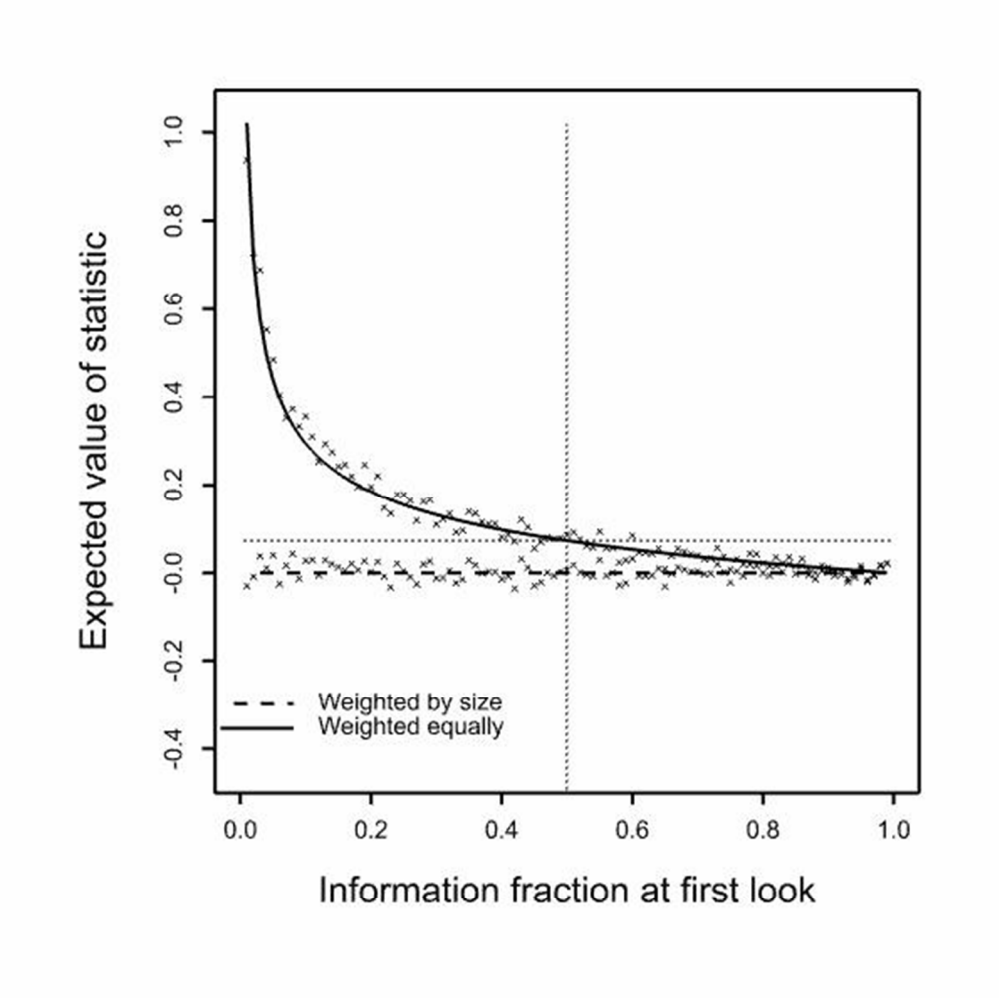
Examination of these issues is beyond this note. The field, however, is an example of the value of considering statistical inference from different perspectives[17, 18].

Acknowledgements

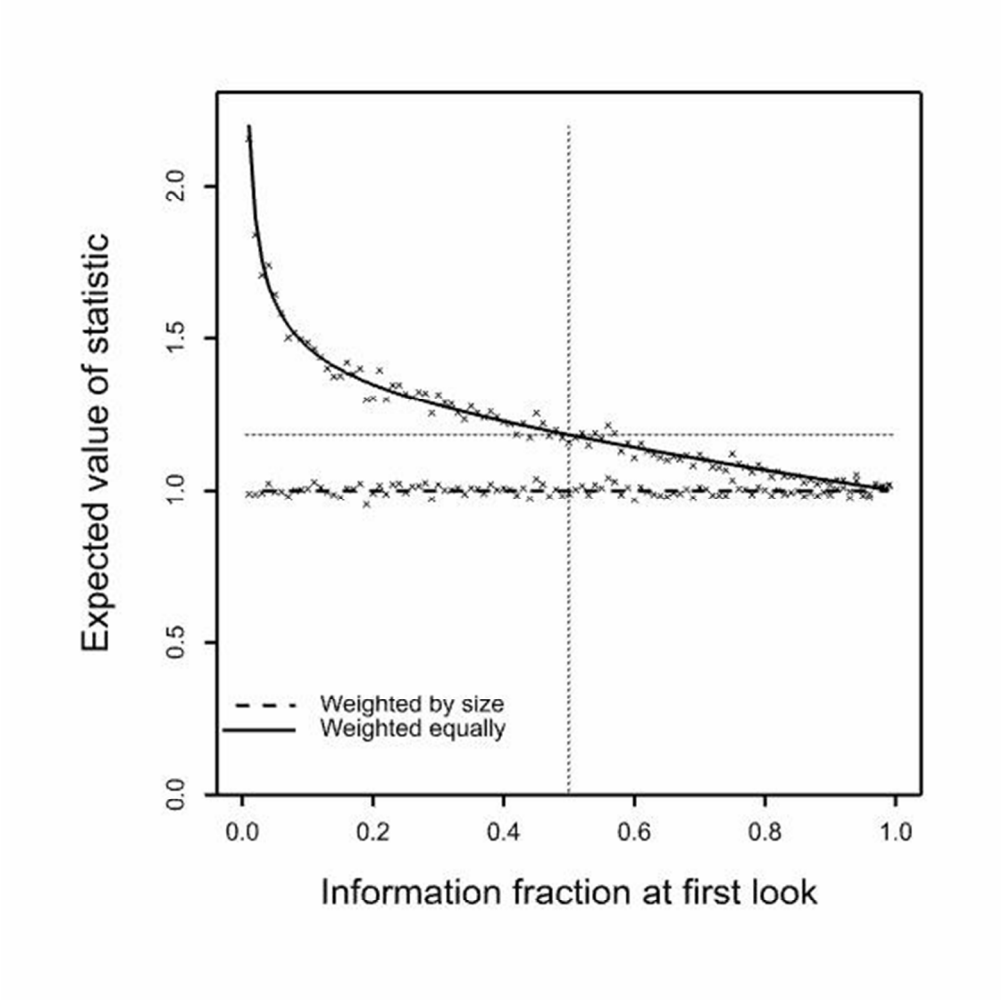
I am grateful to James Roger for illuminating discussions on this issue and to John Whitehead, Stephen Walter and two referees and an editor for helpful comments on earlier drafts. This work has received funding from the European Union's 7th Framework Programme for research, technological development and demonstration under Grant Agreement no 602552.

References

1. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology* 2008; **19**: 640-648.
2. Hughes MD, Freedman LS, Pocock SJ. The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics* 1992; **48**: 41-53.
3. Todd S. Incorporation of sequential trials into a fixed effects meta-analysis. *Statistics in Medicine* 1997; **16**: 2915-2925.
4. Schou IM, Marschner IC. Meta-analysis of clinical trials with early stopping: an investigation of potential bias. *Statistics in Medicine* 2013; **32**: 4859-4874.
5. Royall R. *Statistical Evidence: A likelihood paradigm*. Chapman & Hall/CRC: Boca Raton, 1999.
6. Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L, Gluud LL, Als-Nielsen B, Gluud C. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *Int J Epidemiol* 2009; **38**: 276-286.
7. Pogue JM, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Control Clin Trials* 1997; **18**: 580-593; discussion 661-586.
8. van der Tweel I, Bollen C. Sequential meta-analysis: an efficient decision-making tool. *Clinical Trials* 2010; **7**: 136-146.
9. Higgins JP, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. *Statistics in Medicine* 2011; **30**: 903-921.
10. Senn S, Holford N, Hockey H. The ghosts of departed quantities: approaches to dealing with observations below the limit of quantitation. *Stat Med* 2012; **31**: 4280-4295.
11. Senn S. Transposed conditionals, shrinkage, and direct and indirect unbiasedness. *Epidemiology* 2008; **19**: 652-654; discussion 657-658.
12. Senn S. Being Efficient About Efficacy Estimation. *Statistics in Biopharmaceutical Research* 2013; **5**: 204-210.
13. Burman CF, Sonesson C. Are flexible designs sound? *Biometrics* 2006; **62**: 664-669; discussion 670-683.
14. Senn SJ. The many modes of meta. *Drug Information Journal* 2000; **34**: 535-549.
15. Gessa D. Simple strategies for variance uncertainty in meta-analysis. In *Department of Statistics*. University of Glasgow: Glasgow, 2008.
16. Senn S. Hans van Houwelingen and the Art of Summing up. *Biometrical Journal* 2010; **52**: 1-10.
17. Barnard GA. Fragments of a statistical autobiography. *Student* 1996; **1**: 257-268.
18. Senn SJ. You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets and Morals* 2011; **2**: 48-66.



Two approaches to combining information from sequential trials. It is assumed that there is one look, the true treatment effect is zero, the standard error at full information is 1 and that no adjustment for stopping is made. Theoretical and simulated means (based on 4000 realisations) are shown. Weighted by size: dashed line and crosses. Un-weighted: solid line and circles. An information fraction of 0.5 is highlighted using a dotted line.
157x157mm (96 x 96 DPI)



Two approaches to combining information from sequential trials. It is assumed that there is one look, the true treatment effect is one, the standard error at full information is 1 and that no adjustment for stopping is made. Theoretical and simulated means (based on 4000 realisations) are shown. Weighted by size: dashed line and crosses. Un-weighted: solid line and circles. An information fraction of 0.5 is highlighted using a dotted line.
157x157mm (96 x 96 DPI)

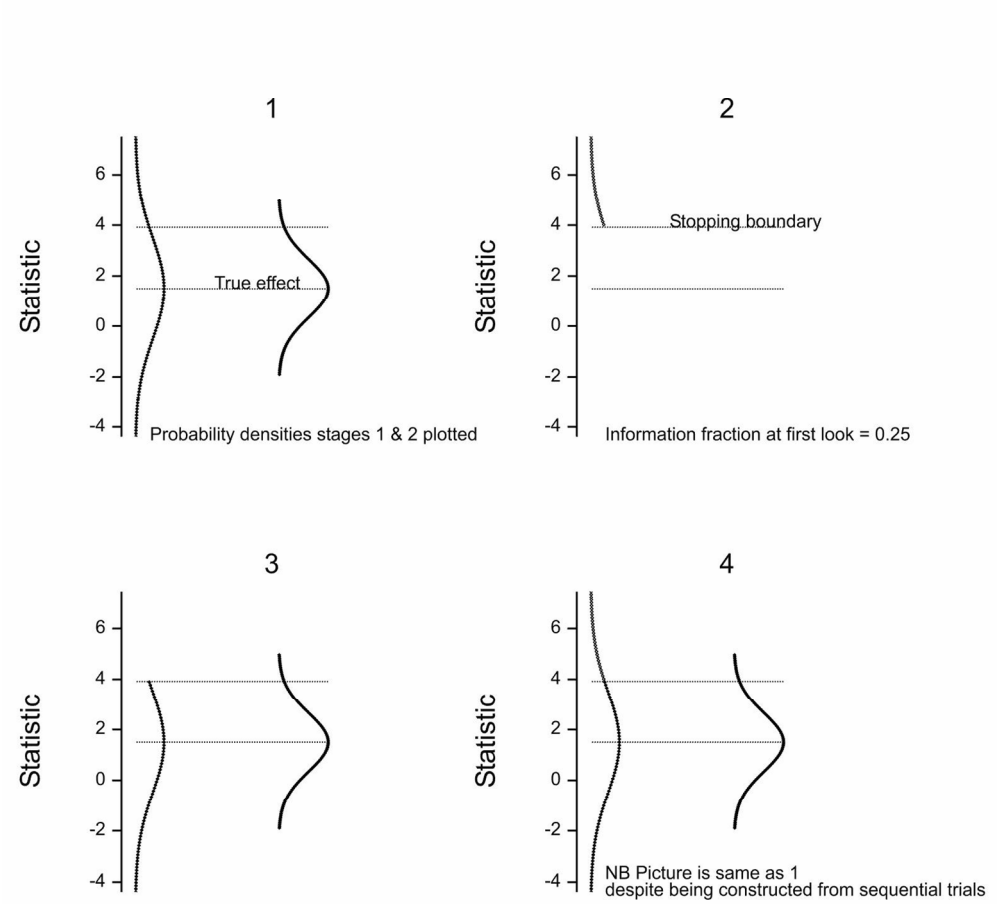


Diagram illustrating combination of information from a sequential trial with one interim look at ¼ of the information. 1. Probability density functions (pdf) for stage 1 & 2 assuming no stopping. 2. Fraction of pdf if stopped early. 3. Fraction of pdf for stage 1 and full pdf for stage 2 if continued. 4 Combined pdfs from 2 & 3.

423x423mm (96 x 96 DPI)