Published in final edited form as:

Res Synth Methods. 2013 September ; 4(3): . doi:10.1002/jrsm.1088.

A Pocock approach to sequential meta-analysis of clinical trials

Jonathan J Shuster, Ph.D.^{1,*} and Josef Neu, M.D.²

¹Department of Health Outcomes and Policy, College of Medicine, University of Florida, Gainesville, FL 32610

²Department of Pediatrics, College of Medicine, University of Florida, Gainesville, FL 32610

SUMMARY

Three recent papers have provided sequential methods for meta-analysis of two-treatment randomized clinical trials. This paper provides an alternate approach that has three desirable features. First, when carried out prospectively (i.e. we only have the results up to the time of our current analysis), we do not require knowledge of the information fraction (the fraction of the total information that is available at each analysis). Second, the methods work even if the expected values of the effect sizes vary from study to study. Finally, our methods have easily interpretable metrics that make sense under changing effect sizes. While the other published methods can be adapted to be “group sequential” (recommended), meaning that a set number and timing of looks are specified, rather than looking after every trial, ours can be used in both a continuous or group sequential manner. We provide an example on the role of probiotics in preventing necrotizing enterocolitis in preterm infants.

Keywords

Bayes; Group Sequential; Meta-analysis; Pocock; Relative risk; Sequential analysis

1. INTRODUCTION

While most meta-analysis applications do not need to employ sequential methods, for a substantial minority, use of such methods is critically important. In drug and device development, it often takes a large series of human trials to properly evaluate a product. As results get published, any individual can assemble the ongoing results into a meta-analysis. The company faces multiple jeopardy in both time and dimension as different side effect analyses can be done. Unlike clinical trials, there is no complete registry of meta-analyses. Selective reporting is very plausible, where an analysis is done and submitted or accepted for publication only if significant. The Cochrane Collaborative, though designed in part to prevent this, is not a complete answer as selective registration is plausible. As the best defense against selective publication of meta-analysis, the company should plan for a sequential meta-analysis from the initiation of randomized Phase II trials of their product.

In order to conduct a rigorous prospective sequential meta-analysis, the following essential qualities of the analysis should be considered as highly desirable. (1) The method should have the ability to allow for target population effect size changes with time. Effect sizes may change over time due to changing follow-up, eligibility, controlling known side effects, etc. with increasing calendar time. (2) The methods should have, at least to a good approximation, independent increments of information about effect sizes from analysis to

*Correspondence to: Jonathan J. Shuster, PO Box 100177, Gainesville, FL 32610-0177. jshuster@biostat.ufl.edu.

analysis. Methods relying on inverse variance weights cannot be reliably used in this context, as relative reweighting of prior studies destroys this important property. Apart from methods relying on Bonferroni style corrections, all classical sequential methods rely on independent increments of information. (3) The method should be robust against incorrect specification of the information fraction (ratio of the amount of information available at the present analysis to the total information at the final analysis that would occur if no stopping barrier was exceeded). This fraction is never known at the time of the interim analysis. (4) Methods should have a physically interpretable effect size. As discussed below, other than methods with conservative Bonferroni style corrections, no published method can meet all of these needs, and hence we propose new methods in this paper that satisfy all four requirements.

Sequential methods are also very useful in assessing a series of trials retrospectively. For example, one may want to second guess whether or not there were excessive trials conducted. When should the answer have been known? For these, fairness dictates that one does not use information beyond the potential stopping point to make this call. Hence, it should be off limits to use as an information fraction the ratio of the information at the look in question to the total information collected after all trials. Our proposed methods do not have this issue. In addition, this capability is extremely valuable to reviewers of meta-analysis papers. The publication of a meta-analysis can adversely affect ongoing trials. For example, Nissen-Wolski (2007) had a very large impact upon the RECORD trial, see Home et al. (2009). Had the reviewers the capability of conducting a sequential meta-analysis, they could have determined whether or not this carefully watched sequence of trials had an effect size that stood up. Without the error control provided by a sequential analysis, repeated analyses of the same question over time can easily lead to spurious associations.

Since 2008, there have been three major publications on sequential meta-analysis of two-treatment randomized clinical trials. Wetterslev et al. (2008) proposed “Trials Sequential Analysis” [TSA], Van der Tweel and Bollen (2010) proposed “Sequential Meta-Analysis” [SMA], and Higgins et al. (2011) proposed a Bayes approach. None of these methods allow for the effect sizes to be dynamic. Random effects are drawn from the same conceptual urn from trial to trial. All of these competitors to our methods reweight the relative contributions of the included trials after each trial is added. This violates the critical independent increment property. A potential shortcoming of all methods (including ours) lies in a lack of knowledge of the true information fraction (the ratio of the variance of the estimate at the final look presuming no stopping to that after the current look). Used prospectively, all of the published methods can actually see an information fraction that goes backward in time as a trial is added! This will be illustrated in the discussion. We chose to use the Pocock group sequential approach for the analysis, based upon the weighted methods of Shuster (2010), for all meta-analyses other than low event-rate binomial trials and upon the weighted method of Shuster et al. (2012) for low event-rate binomial trials. This method has all four properties cited above. The Pocock approach is fairly aggressive toward early stopping when compared to others. It is based on a common marginal P-value at every look. That is, if you ignore the sequential nature of the looks, the P-value required for stoppage is the same at each look. Most sequential methods require a lower P-value at early looks than later looks to stop. But In Section 4, the Pocock method will be seen as highly robust against lack of knowledge of the true information fraction. Even though our tables will accommodate up to 12 looks, we recommend a ceiling on the maximum number of looks taken as up to six. The reason for this is that the greater the number of looks, the lower the P-value needs to be for each look. Further, we need to construct conservative confidence intervals as the effect sizes can change from trial to trial, so methods for repeated confidence intervals cannot be borrowed from group sequential methods for clinical trials. Pure sequential monitoring for

example will tend to have much wider confidence intervals than group sequential monitoring, and can easily lead to later stopping than group sequential methods.

While a sequential approach, which requires very convincing evidence at each look to stop, may seem to be overly conservative, there are very good arguments in favor of it, as we conclude from four articles by Ioannidis. In a retrospective analysis of the time from completion of follow-up to publication vs. the significance of the result, Ioannidis (1998) showed a strong correlation in that non-significant trials had a longer time window (difference in medians of 1.3 years) from completion of follow-up to publication. Ioannidis (2005a) studied 49 of the most highly cited articles of single trials, and of the 45 reporting significant results, 7 (16%) were contradicted by later studies and 7 others (16%) found the effect sizes weaker in subsequent studies. In an essay by Ioannidis (2005b), an effective argument that a substantial proportion of research claims are false is made and supported by simulations. Finally, Ioannidis (2010) argues that despite the strengths of meta-analysis, retrospective meta-analysis has serious difficulties in terms of reporting bias. A key element of that paper, in terms of a single meta-analysis of a question, lies in his discussion of the effects of time, especially that they often diminish over time. He strongly argues for a prospective approach, as Shuster and Gieser (1996) did. We argue that to support the ideas of Ioannidis, the methods need to control errors for multiple looks, as sequential methods do, in order to offer strong protection against false positives. They also need to be flexible in terms of allowing the true effect sizes from trial to trial to change over time. Our methods meet both of these desirable standards.

Section 2.1 is devoted to comparison of means for studies that do not involve rare events. Section 2.2 is devoted to binomial studies with rare events. Section 2.3 describes other applications for adapting Section 2.1 to log odds ratios for non-rare event binomial studies and adapting Section 2.2 to actuarial survival studies. Section 3 presents an example of Deshpande et al (2010) on the use of probiotics to prevent necrotizing enterocolitis. Section 4 does extensive simulations on the robustness of the Pocock approach to uneven accrual of information between analyses, while Section 5 presents a concluding discussion. Appendix 1 is devoted to a non-technical review of the methods.

2. WEIGHTED ESTIMATION OF A VECTOR OF MEANS

We shall adopt methods that borrow from Bayes ideas. We shall impose a separate prior distribution on the effect sizes for each trial. We make no parametric assumption about the nature of the prior distribution, except that it has a mean effect size (possibly multivariate). This is a generalization of the ideas of random effects, since if one assumes the prior distributions are the same for all trials, we have a classical random effects problem. Allowing the prior distributions to be different gives us the needed flexibility to have systematic changes in effect sizes from trial to trial. This distinction is not relevant to non-sequential methods. The sample sizes for each trial are presumed fixed (non-random), but can vary from trial to trial. We employ the following conceptual model for effects, assuming no more than the independence of our studies. For each study j , $j=1,2,\dots$ the estimated univariate or multivariate effect size $\hat{\theta}_j$ satisfies the model:

$$\hat{\theta}_j = \theta_j + \varepsilon_j \quad (1)$$

with the effects, θ_j , drawn at random, from their own conceptual sequential study urns, with

$$E(\theta_j) = \phi_j \text{ and } E(\varepsilon_j) = 0$$

The targeted parameter after K trials will be the patient weighted mean ϕ_j for the first K trials (a moving target unless the ϕ_j are all the same):

$$\Delta_K = \sum N_j \phi_j / \sum N_j \quad (2)$$

with N_j = total number of subjects on Trial j. Δ_K is the patient weighted target population mean vector through the first K trials.

2.1 Differences in Means

Shuster (2010, Section 3) provides methods for non-sequential estimation of patient-level differences in target population means in random effects meta-analysis. For the j-th study, denote the difference in sample means by

$$D_j = (\bar{Y}_{2j} - \bar{Y}_{1j}) \quad (3)$$

and let the total sample size be represented as the sum of the two treatment-specific sample sizes N_{1j} and N_{2j} :

$$N_j = (N_{1j} + N_{2j}) \quad (4)$$

The weighted estimated effect size for the first K trials is:

$$\hat{\Delta}_K = (\sum D_j N_j / \sum N_j) = (\sum U_j / \sum N_j) \quad (5)$$

with $U_j = D_j N_j$.

The physical interpretations of what $\hat{\Delta}_K$ is estimating, namely $E(\hat{\Delta}_K) = \Delta_K$, along with that of its relative risk counterpart for rare event binomial trials will be deferred to Section 5.

Using the delta method of Serfling (1980), Shuster (2010) derived the asymptotic variance of $\hat{\Delta}_K$ for large K as:

$$V_K = [\{ \text{Var}(U_j) / E^2(N_j) \} + \{ \text{Var}(N_j) E^2(U_j) / E^4(N_j) \} - 2 \{ \text{Cov}(U_j, N_j) E(U_j) / E^3(N_j) \}] / K \quad (6)$$

E=Expected value, Var=Variance, and Cov=Covariance in (6).

This variance was derived by Shuster (2010), by noting that after a random rearrangement of the indices 1,2,...,K the value of the estimate in (5) is unchanged. After such a permutation, the K pairs (U_j, N_j) are identically distributed. The permutation makes the N_j random.

A consistent estimator for V_K , $\hat{V}_K$, is obtained by substituting the sample means, variances, and covariance into (4): for $U_j = D_j N_j$ and N_j . Our sample moments use K-1 as their denominators (not K).

We **define a set of trial numbers**, Ω , where looks can take place. For example if $\Omega = \{6, 11, 14\}$ we will look only after trials 6, 11, and 14 in our group sequential plan.

After trial K, we define the normalized statistic as

$$T_K = \hat{\Delta}_K / \text{SQRT}(\hat{V}_K) \quad (7)$$

The cutoff is determined entirely by the overall significance level α , the number of looks (L) and the trial number (K).

We shall define the cutoff as $T(\alpha, L, K)$, as

$$T(\alpha, L, K) = -\text{TINV}(\gamma/2, K-2) \quad (8)$$

where TINV (readily available in SAS and other packages) is the inverse of the central t-cumulative distribution with $K-2$ degrees of freedom and γ is the nominal level obtained for the Pocock cutoff in Table 1. The use of $K-2$ degrees of freedom is motivated by Shuster et al. (2012). We produced Table 1 through simulation, with 1,000,000 replications under the classical Pocock (1977) model, with a maximum number of looks ranging from 2–12. While Pocock used a constant Z-value as a cutoff, it is more common today to use a constant marginal (non-sequential) P-value, denoted by γ in our tables.

For example, if $\alpha=0.05$, $K=10$ studies completed so far, and $L=4$ (maximum of four looks), $\gamma=0.0184$, which yields $T(\alpha, L, K)=2.95$.

For example, if $\alpha=0.05$, $K=10$ studies completed so far, and $L=4$ (maximum of four looks), $\gamma=0.0184$, which yields $T(\alpha, L, K) = -\text{TINV}(0.0092, 10-2) = 2.95$.

The continuation/stopping rule is defined below:

- i. **Continue** doing studies if the current trial number is not in the set Ω and the process has not been halted.
- ii. **Stop** doing studies if either $K = \text{Max}\{\Omega\}$ or K belongs to Ω and $|T_K| > T(\alpha, L, K)$

Sequential Properties—First note that even if $\Delta_K = E(\hat{\Delta}_K)$ is changing with K , and presuming for the moment that we will force completion of all studies irrespective of their results,

$$P\{|\hat{\Delta}_K - \Delta_K| < |\text{TINV}(\gamma/2, K-2)| \text{SQRT}(\hat{V}_K) \text{ for all } K \in \Omega\} \approx 1 - \alpha. \quad (9)$$

Hence, regardless of the stopping point, any event that contains this intersection in (9) must have at least as high a coverage probability, and hence we have more than $100(1-\alpha)\%$ confidence that

$$|\hat{\Delta}_M - \Delta_M| < |\text{TINV}(\gamma/2, K-2)| \text{SQRT}(\hat{V}_M) \quad \text{where } M = \text{actual stopping time.} \quad (10)$$

This forms a conservative (naïve) $100(1-\alpha)\%$ confidence interval for Δ_M .

2.2 Rare Event Binomial Trials

In this section, we provide a direct analog to the previous subsection using Shuster et al (2012), Section 4 instead of the methods of Section 2.1.

We define the patient level relative risk Treatment 2: Treatment 1 as

$$RR_K = \sum N_j P_{2j} / \sum N_j P_{1j} \quad (11)$$

where N_j is the total sample size of the j -th trial (regardless of allocation percentage), $j=1,2,\dots$ and P_{ij} are the true target population failure rates for the i -th treatment ($i=1,2$) and j -th trial.

Let $\bar{A}_{iK}$ represent the sample mean of the first K A_{ij} values for treatments $i=1$ and $i=2$, respectively, where $A_{ij} = N_j \hat{P}_{ij}$ and $\hat{P}_{ij}$ is the observed failure rate on Treatment i , Study j , $j=1,\dots,K$, respectively.

After studies $1,2,\dots,K$, we estimate the relative risk RR_K , as the ratio of the sample means of the A_{ij} :

$$\widehat{RR}_K = \bar{A}_{2K} / \bar{A}_{1K} \quad (12)$$

From Shuster et al (2012) as $K \rightarrow \infty$, the normalized asymptotic distribution of $\log(\widehat{RR}_K)$ is Student central T with $(K-2)$ df, asymptotic mean $\log(RR_K)$ and variance

$$V_K = [\text{Var}(A_{1j})/E(A_{1j})^2 + \text{Var}(A_{2j})/E(A_{2j})^2 - 2\text{Cov}(A_{1j}, A_{2j})/\{E(A_{1j})E(A_{2j})\}]/K \quad (13)$$

By substituting the first and second sample moments for the population moments of the A_{ij} ($K-1$ denominator) into (13), we obtain a consistent estimator for V_K , $\widehat{V}_K$. From this point, we use $\log(\widehat{RR}_K)$ and $\log(RR_K)$ as the analogs of $\hat{\Delta}_K$ and Δ_K respectively from the previous section, and make inferences in the same way. For confidence intervals in the original scale, we must take natural antilogs of the point estimate and endpoints of the interval.

2.3 Other Applications

Binomial studies in which the events are not rare—The roles of the D_j can be played by the study specific log odds ratio estimate in Section 2.1.

Actuarial survival studies—This could be done per 2.2, with the role of the sample sizes being played by the total time on test. This yields estimates of mean hazard ratios. Unfortunately, the total time on test is not always available in the published studies.

3. NUMERICAL EXAMPLE

Deshpande et al (2010) combined eleven randomized trials pitting probiotics against a control for preterm infants to study three outcomes (necrotizing enterocolitis [NEC], sepsis, and mortality) with NEC the primary variable. These are **low event rate** binomial trials, and hence, we shall employ the methods of Section 2.2. We shall restrict our analysis to NEC. The Deshpande (2010) analysis used a weighted sequential analysis, which not only used an empirical weighting method for low events against the advice of DerSimonian-Laird (1986) and the Cochrane Handbook (2008 and 2011), but it also does not meet the criteria we proposed at the start of the Introduction. Note that we are aware of additional randomized trials, three ongoing and one about to start. We therefore reasonably plan for a last look at 15 trials. Suppose that the long term strategy was to conduct a group sequential analysis with plans to look after the 5th, 9th, 12th, and 15th trial. The results of our analysis were obtained using the calculations from (12) and (13), and substituting these into (9). As indicated in Table 2, we see that the group sequential analysis would be halted after nine trials, with a

point estimate for the relative risk of 0.32 (95% confidence interval from 0.16–0.65). The actual Trial Sequential Analysis run by Deshpande et al, Figure 6, coincidentally also crossed its stopping barrier at nine trials, based on a fixed effects analysis. Note that our confidence interval not only excludes the neutral value of relative risk of 1.00, but it includes only clinically significant values, since a relative risk of 0.65 represents a 35% decrease in risk.

How the confidence interval for the relative risk, RR at the final look is constructed: We obtained the point estimate for the relative RR, after trial 9, from Table 2 as $\exp(-1.14)=0.32$. From Table 1, we see that for a maximum of four looks, and overall significance of 0.05, $\gamma=0.0184$.

$|TINV(\gamma/2, 9-2)| = T\text{-cutoff for } \gamma/2=0.0092 \text{ and } 7 \text{ df}=3.05$. The Standard error in the log scale from Table 2 is 0.23. The lower and upper limits of the confidence interval are respectively

$$LL=\exp(-1.14-3.05*0.23)=0.16 \text{ and } UL=\exp(-1.14+3.05*0.23)=0.65$$

As an exercise, suppose we ran a continuous sequential analysis with planned looks after each trial starting from the 4th and ending at the 15th (12 planned looks). Had this been utilized with our Pocock bounds, we would have also stopped at the 9th trial with a point estimate for relative risk of 0.32 and 95% confidence interval from 0.14 to 0.71. However, the series would have almost terminated on the 8th trial, which had its confidence interval from 0.11 to 1.001. Had the stop occurred with the eighth trial, the upper limit would have been a clinically insignificant effect. Certainly it would have been a favorable outcome to stop earlier, but continuous monitoring will often lead to confidence intervals that include at least some clinically insignificant results. They also could fail to ever stop early as increasing the number of looks increases the level of evidence required at each look. Deshpande et al. (2010) reported a point estimate of 0.35 for relative risk, and a non-sequential 95% confidence interval of 0.23 to 0.55. This confidence interval, unlike ours, does not take multiple looks into account. This was confirmed on commercial software.

4. SIMULATION STUDY

In this section, we look at a key question for application of the Pocock (1975) method. It was designed for equal increments of information, whereas we expect the increments to be quite variable. The question addressed in this section is whether the method is sufficiently robust against departures from equal increments of information to make it a reliable tool. If robustness is not sufficient, a secondary question is whether or not the robustness will hold under restricted conditions. Accordingly, we developed a large simulation using normally distributed observations and random increments of relative information, for maximum number of looks ranging from 2 to 6.

For relative information, we generated uniform random variables over the interval 0.1 to 0.9. The total information was proportional to the relative information. For example if we were working with a maximum of three looks, and the random numbers were 0.15, 0.85, and 0.45, the information for Total=0.15+0.85+0.45=1.45, would be 0.15/1.45= 10.3% after the first look, (0.15+0.85)/1.45=69.0% after two looks, and 100% after three looks. This is quite different from equal increments of 33.3%. 66.7% and 100%. The idea of selecting relative information was to have each increment being substantial but allowing major departures from uniformity. We also looked at two other constrained experiments, increasing increments (decreasing increments) of information over time, where after selecting the relative information as above, we sorted them in increasing (decreasing) order respectively.

Before conducting the experiments, we expected that random increments would prove to be quite accurate, increasing increments to be anti-conservative- at least some extent, and decreasing increments to be conservative- at least to some extent. We expected that the later the looks relative to equally spaced, the higher the coverage would be. For each of these experiments we generated 10,000 randomly selected scenarios described above, and simulated the data needed for the Pocock analysis 100,000 times per scenario. Table 3 provides the performance of the simulations.

As can be seen in Table 3, indeed if the increments of information were forced to be decreasing in time, the coverage was never below 94.7% (predicted to be 95% for equal spacing), but looking at the 90th percentiles, it becomes somewhat conservative when one has five or six looks. For totally random selection of incremental information, the coverage was quite accurate, with the interval from the 10th to the 90th percentiles within 1% of the targeted 95%. For increasing increments, the coverage, as predicted, was somewhat anti-conservative. But since the interval in (10) created by the actual stopping time already has more coverage probability than the intersection of the intervals described in (9), we conclude that even for increasing intervals, the Pocock method should be considered reasonably accurate in terms of the coverage described in (10). Nonetheless, it is prudent to space looks by first looking after a reasonably high percentage of the trials have been completed, especially since later trials tend to be larger than earlier ones.

5. DISCUSSION

How to interpret the effect size in a dynamic environment

We consider this natural pseudo-experiment, thought of before the effect sizes are drawn randomly from their conceptual study-specific urns and before the first subject is treated. Pick an unassigned subject at random from the first K trials and an effect size from the conceptual study-specific urn of effect sizes for the selected study. What is the metric for the subject being assigned to Treatment #2 to that if assigned to Treatment #1? For example, for mean differences, this is simply the difference in expected responses. For relative risk, it is the ratio of failure rates Treatment #2: Treatment #1. Note that the targeted metric can change in value with changing values of K . Note that the only other weighting system that can provide a parallel physical interpretation of the effect size metric is an unweighted analysis, where the unit of random selection is the study rather than the subject. With dynamically changing effect sizes, optimization of weights is no longer relevant, since each weighting system is targeting a different overall effect size parameter. Any other weighting scheme beyond patient or study based will not be accompanied by a physically defined effect size interpretation.

Caution

Our inference from the Pocock method applies only to the trials at or before the time when the actual stop occurs. Unlike sequential methods for single trials, in which a stopping rule such as the O'Brien-Fleming (1979) method is used, we cannot claim that the conclusion is unlikely to change if only we completed the entire study sequence. Unknown creep toward the null might occur in the future in a meta-analysis scenario that is much less likely in a single trial. This issue applies to any sequential meta-analysis method, however.

The methods of Wetterslev et al. (2008) and Higgins et al. (2011) both require independent increments of information plus the assumption that there is no "creep" in K of the true effect sizes after K trials. They reweight the studies after each study is added, violating the independent increment premise and can even see a decrease in the total information as a trial is added, in direct violation of what is supposed to occur in standard sequential statistical methods. For a concrete example, consider the scenario in Table 4 where for simplicity we

know all of the within and between study variances, and therefore can appropriately weight the estimates at each stage as proportional to the inverse of the variances. The true variance of the estimate after three trials is actually smaller than that after four trials, indicating that the information time has indeed moved backward. Second, from the changing relative weights of the first three studies with vs. without the fourth study, the added information from study 3 to study 4 is not independent of that from the first three trials. To achieve independent increments of information as required by the underlying sequential methods used in Wetterslev et al. (2008) or in Higgins et al. (2011), the estimate after the fourth trial should have been, at least approximately, a linear combination of the estimate after three trials and the estimate for the fourth trial. Therefore, these sequential methodologies do not in general meet the basic conditions required by the underlying methods they are employing. Our methods employ weights proportional to the total sample size, and therefore keep the relative weights (e.g. ratio of weight of trial 1 to weight of trial 2) the same as we add trials. For inverse variance-based weights, this type of adverse behavior is likely when the first set of trials are relatively homogeneous and the next set of trials have members at odds with the first set.

Sequential Meta Analysis, as described by van der Tweel and Bollen (2010) has fewer issues than Wetterslev et al. (2008) or Higgins et al. (2011). It is a close analog of unweighted methods. Each trial contributes a Z-score (standardized test statistic for no effect), and these Z-scores are accumulated unweighted as trials are added. The accumulating information is the sum of the estimated variances for these studies. Under fixed effects, the only requirement would be knowledge of the total information given there will be no stopping. However, if there are random effects, then the between study variances must be taken into account for the estimation of the study-specific variance in the Z-statistic, rendering it as having some of the same problems, such as non-independence of increments and going backward in information time, as the other two published methods. Creep in the effect sizes over time will wreak further havoc upon this method. This method can be applied retrospectively, however, once all the trials are in, and a second guess as to a stopping time is desired.

There is a closely related paper by Whitehead (1997) that tackles a slightly different problem. This deals with a series of similar trials that are running concurrently, and the analyst has access to the interim data at various points in time. From a design point of view, this approach is very attractive. However, from a practical point of view, the availability of such data sets would be limited to governments and the pharmaceutical companies. When employed prospectively using empirical weighting at each stage, and if new trials can be initiated between stages, there will be the same issues of independence of information increments and true effect size changes that plague prospective use of Wetterslev et al. (2008), van der Tweel and Bollen (TSA, 2010), and Higgins et al. (2011). If there are truly fixed effects, the Whitehead (1997) method seems to be applicable, however.

Although we used our **weighted** methods as described in Shuster (2010 Section 3) and Shuster et al (2012 Section 4), we could equally well have used **unweighted** methods per Shuster (2010 Section 2) and Shuster et al (2012 Section 3,) the latter in both cases for low event-rate binomial trials. The choice as to the approach should be made at the planning stage of the series of trials. When indeed there is no association between effect size and sample size, the two approaches will estimate the same target parameter. If this is the case, at least in the low event binomial non-sequential random effects scenario, the weighted method (using t with $df=K-2$, K =# of studies to date in this analysis) was superior to the unweighted (t cutoff $df=K-1$) for the length of the confidence interval, but inferior in terms of closeness to the desired 95% coverage. Normal distribution cutoffs, unlike the t -cutoffs, were not accurate.

As we noted in Section 3, the use of a group sequential approach will tend to be associated with somewhat later stopping than a continuous sequential approach. The tradeoff is that it is far more likely that when a continuous sequential analysis is stopped for significance, the resulting sequential confidence interval will include clinically insignificant outcomes, as compared to a similar happening for a group sequential approach.

Flexibility of Timing

As pointed out by a colleague (John Kairalla, Assistant Professor of Biostatistics, University of Florida), the robustness of the Pocock method means that the planned timing of analysis could be based on calendar time, rather than upon trial number. For example, we could plan for the analysis of trials completed every three years for 12 years (4 looks), rather than to look after a pre-specified number of trials were completed (e.g. 6, 9, 12, 15). In addition, if studies on the intervention are discontinued for reasons not related to the outcome, this calendar approach allows one to conduct a final analysis with fewer studies than hoped for. For example, if we planned to have four looks, but after two looks, the studies were discontinued, (10) is still a valid conservative confidence interval for its effect size. Consistent with intent-to-treat principles, the total looks in (10) should remain at four and not be reduced to two.

Concluding Remarks

The Pocock approach, under carefully planned increments of studies, offers a robust method for conducting prospective group sequential meta-analysis of a series of clinical trials. For pharmaceutical companies, it would be a good idea to set up independent oversight boards for each product, starting from the first human research. This protects manufacturers from multiple jeopardy from unstructured repeated analyses and multiple endpoints. The methods can be easily adapted to multi-dimensions, by using α/P instead of α for each of P dimensions. We have provided cut points for $P=2,3,4$, and 5 in Table 1. For example, if $P=3$, we would use $\alpha=0.05/3=0.0167$ instead of 0.05. Looks for different dimensions need not all be conducted at the same point in time, nor do they require the same number of planned looks. Finally, if one uses sequential methods, interval estimates should take multiple looks into account (as we do), rather than reporting intervals calculated as if the analysis was not sequential in nature. As suggested by a reviewer, there is scope for a parametric Bayes analog to our methods, and we encourage other researchers to solve this problem. A paper that offers great promise to develop this approach is due to Schmid and Brown (2000).

A SAS macro for calculating the results for the analysis of Section 3 and Table 2 is available on the “Jon Shuster’s SAS design and analysis programs” button on <http://ags.bwh.harvard.edu/>, the website for Clinical and Translational Science Statisticians.

Acknowledgments

This work was partially supported by NIH grants 1UL1TR000064 from the National Center for Advancing Translational Sciences and R01 HD059143 from the National Institute of Child Health and Human Development. The authors wish to thank Dr. Ingram Olkin, Stanford University who reviewed an early draft and provided helpful suggestions. Thanks also go to Dr. John Kairalla, University of Florida, for his helpful comments on flexible timing. In addition, we appreciate the helpful comments of the editor, associate editor, and reviewers, whose contributions had a major impact upon both the theory and the presentation of the material.

References

- DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986; 7(3):177–188. [PubMed: 3802833]

- Deshpande G, Rao S, Patole S, Bulsara M. Updated Meta-analysis of Probiotics for Preventing Necrotizing Enterocolitis in Preterm Neonates. *Pediatrics*. 2010 May; 125(5):921–930. [PubMed: 20403939]
- Higgins, JPT.; Green, S. *Cochrane handbook for systematic reviews of interventions*. Wiley-Blackwell; Chichester UK: 2008, 2011.
- Higgins JPT, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. *Statistics in Medicine*. 2011; 30:903–921. [PubMed: 21472757]
- Home PD, Pocock SJ, Beck-Nielsen H, et al. RECORD Study Team. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre randomised, open-label trial. *Lancet*. 2009; 373:2125–2135. [PubMed: 19501900]
- Ioannidis JPA. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*. 1998; 279 (4):281–286. [PubMed: 9450711]
- Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005a; 294 (2):218–22. [PubMed: 16014596]
- Ioannidis JPA. Why most published research findings are false. *PLOS Medicine*. 2005b; 2 (8):696–701.
- Ioannidis JPA. Meta-Research: The art of getting it wrong. *Research Synthesis Methods*. 2010; 1(3): 169–184.
- Jennison, C.; Turnbull, BW. *Group sequential methods with applications to clinical trials*. Chapman Hall/CRC Press; Boca Raton FL: 2000. p. 189-204.
- Nissen SE, Wolski K. Effect of Rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med*. 2007 Jun 14; 356(24):2457–71. [PubMed: 17517853]
- O'Brien PC, Fleming RT. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35:549–556. [PubMed: 497341]
- Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*. 1975; 31(1):103–15. [PubMed: 1100130]
- Schmid, CH.; Brown, EN. Bayesian hierarchical models: in *Methods in Enzymology*. In: Johnson, ML.; Brand, L., editors. *Numerical Methods Part C*. Vol. 321. Academic Press; New York: 2000. p. 305-330.
- Serfling, RJ. *Approximation theorems in mathematical statistics*. John Wiley Publication; New York: 1980. p. 118-125.
- Shuster JJ. Empirical vs. natural weighting in random effects meta-analysis. *Statistics in Medicine*. 2010; 29:1259–1265. [PubMed: 19475538]
- Shuster JJ. Review: *Cochrane handbook for systematic reviews for interventions*. *Research Synthesis Methods*. 2011; 2:126–130.
- Shuster JJ, Gieser P. Meta-Analysis and Prospective Meta-Analysis in Childhood Leukemia Clinical Research. *Annals of Oncology*. 1996; 7:1009–1014. [PubMed: 9037358]
- Shuster JJ, Guo JD, Skylar JS. Meta-analysis of safety for low event-rate binomial trials. *Research Synthesis Methods*. 2012; 3(1):20–30.
- Van der Tweel I, Bollen C. Sequential meta-analysis: an efficient decision-making tool. *Clinical Trials*. 2010 Apr; 7(2):136–46. [PubMed: 20338906]
- Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *J Clin Epidemiol*. 2008 Jan; 61(1):64–75. [PubMed: 18083463]
- Whitehead AA. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in Medicine*. 1997; 16:2901–2913. [PubMed: 9483722]

APPENDIX 1: NON-TECHNICAL SUPPLEMENT

This section describes the content and the motivation behind using the methods described in this paper. Imagine that you are developing a pharmaceutical product aimed at treating a specific disease. As results of your trials are serially published, any meta-analyst can look at the outcomes and side effects and perform their own analysis. It is likely they will only

publish an analysis if it shows something significant, and if this is a side effect, your company may be sued for damages. These analyses, even if done in good faith, subject your company to multiple jeopardy in both time (as publications of single trials accumulate) and in dimensions (several different side effects may be under scrutiny). An organized sequential analysis plan offers the best protection against these issues. As an example, just for a single outcome, if one used repeated analysis that treated each look on an individual basis (not accounting for study-wide errors), there is a 14% chance of falsely seeing a “significant difference” at $P < 0.05$, somewhere after the 6th trial in a fixed effects analysis of 15 identical trials where there was actually no effect. If several outcomes are so monitored, the risk of false positives is even higher. When there are a key set of outcomes to monitor, we have provided the necessary ingredients in Table 1 to modify the P-value for the sequential methods to account for this. For example, for two outcomes we would replace 0.05 by 0.025 as our α -level.

The methods we propose for our analysis have several strengths. First, they allow the effect sizes for different studies to be drawn from conceptually different populations. All classical random effects methods presume the effect sizes were drawn randomly from a single population. If there are trends, for example, in the true effect sizes, these classical methods (but not ours) have validity issues. There can be trends in either direction, where the effect sizes migrate toward the null as the number of trials increase (e.g. developments in concomitant care might decrease the effectiveness of the intervention over time) or migrate away from the null (e.g. new ways of mitigating the side effects of the drug, or changes in eligibility over time to weed out unfavorable segments of the target population, might move the effect sizes in this direction over time). For more detail on this issue, see Shuster (2011). Second, by using sample size weights, we obtain a simple physical interpretation of the metric we are estimating. For example, relative risk for the first nine studies is defined as follows: Pick an unassigned patient at random from the first nine trials, and pick an outcome at random from the target population of outcomes in its conceptual urn of outcomes for that trial. What is the ratio of probability that the subject has an adverse event if assigned to Treatment #2 to that if assigned to Treatment #1? Other than an unweighted approach, this type of interpretation is not possible. (Unweighted methods pick a study at random, then an unassigned patient from that study at random). Third, our methods offer simple group sequential confidence intervals for the effect sizes, controlling for overall study errors due to repeated looks.

The key innovation to allow us to evaluate the sequential distributional properties lies in the fact that after a random rearrangement of the study indices 1,2,...,K (a) the estimates and standard errors are unchanged and (b) the couples (sample size, observed mean) all have the same distribution after this random permutation. This enables us to obtain point estimates and standard errors after any number of trials. We have used this concept in Shuster (2010) and Shuster et al. (2012).

On “Independent Increments of Information”

The example in Section 5 shows this can be seriously violated for sequential methods that rely on inverse variance based weights in random effects. We now illustrate why indeed our methods, which use sample size weights, have independent increments, as long as the individual studies are independent per standard assumptions in meta-analysis. The key is to show that if you have two increments of studies, you will obtain exactly the same estimate if you (a) use your weighting system on the pooled set of studies as (b) first use your weights to get the estimates of the studies in each increment and then combine the increment-specific estimates using your weighting system. This will be illustrated for sample size weighted estimates in Table 5. Let us consider studies 1–6 and 7–9 as separate increments of

information for two looks. The estimated effect size for Studies 1–9 is the sample size weighted average of the first nine studies. Mathematically, this is exactly the sample size weighted average of the effect size estimate from first six studies and the effect size estimate from studies 7–9. Table 5 shows the details. If you use inverse variance weighted estimates in random effects methods, you will change the relative weights of studies 1–6 making the increments non-independent, as the example in Section 5 points out.

Independent increments vs. independence of studies are two different concepts. All meta-analysis methods, sequential or not, assume the study results are independent. But if relative reweighting can occur, then independent increment assumptions will be violated.

Trade-offs between continuous sequential methods (look after every trial from a certain point on) and Group Sequential methods (look only after pre-selected numbers of trials are completed)

On the plus side, continuous sequential methods will tend to stop earlier for significance than group sequential methods. But since the stop is just after a study where the confidence interval contained the null value, and the next study usually has a relatively small increment of information, the confidence intervals will tend to exclude the null value by a smaller margin than a group sequential approach. As single studies are added, the confidence intervals tend to move slowly from one to the next. If it is important for the confidence interval to include only clinically important departures from the null value, group sequential approaches have an advantage. This was illustrated in Section 3. Another advantage for Group Sequential methods is that for continuous methods, the addition of a single study represents a small addition, and the large sample approximations may be tenuous, when compared to group sequential methods which might add 3+ studies per increment.

Why the Pocock (1975) group sequential method was chosen

Sequential methods in general require knowledge of the information fraction (the amount of information collected to date divided by the amount of information that would be collected if all of the studies were done, and no early stopping occurred.) This is the anticipated variance of the estimate at the final analysis divided by the variance at the current analysis. This is generally unknown. Because the Pocock method relies on the same fixed P-value at each look, we expected it to be robust against this lack of knowledge. This was confirmed in our simulation.

How to operationalize the Group Sequential Approach

The methods are best applied prospectively. In advance of doing the first study, we need to provide planning values for the timing of the looks, generally recommended to be at decreasing intervals as the number of trials increase. Even if the drug does not make it to where this number of studies gets done, the plan is well documented in case it does. The analysis, based upon only some of the planned looks yields, conservative confidence intervals, even in the absence of completing the entire sequence of trials.

Table 1

Normal Cutoffs for Pocock Group Sequential Procedure. Entries are the cutoff by the Normal distribution (Z), and the nominal two-sided significance level represented by the standard normal distribution. $P(|Z| > 2.18) = 0.0294$. The values in () are used as γ in equations (9) and (10).

Maximum Looks (2-6)						
Level (α)	2	3	4	5	6	
0.05	2.18(0.0294)	2.29(0.0221)	2.36(0.0184)	2.41(0.0159)	2.45(0.0142)	
0.025	2.45(0.0143)	2.55(0.0106)	2.62(0.00875)	2.67(0.00750)	2.71(0.00670)	
0.0167	2.60(0.00945)	2.70(0.00696)	2.77(0.00568)	2.81(0.00489)	2.85(0.00435)	
0.0125	2.69 (0.00706)	2.80(0.00516)	2.86(0.00419)	2.91(0.00362)	2.95(0.00322)	
0.010	2.77(0.00561)	2.87(0.00410)	2.93(0.00335)	2.98(0.00287)	3.02(0.00253)	

Maximum Looks (7-12)						
Level(α)	7	8	9	10	11	12
0.05	2.48(0.0130)	2.51(0.0120)	2.53(0.0113)	2.55(0.0107)	2.57(0.0102)	2.59(0.00972)
0.025	2.74(0.00609)	2.77(0.00563)	2.79(0.00526)	2.81(0.00498)	2.83(0.00471)	2.84(0.00450)
0.0167	2.88(0.00394)	2.91(0.00365)	2.93(0.00342)	2.95(0.00322)	2.96(0.00306)	2.98(0.00292)
0.0125	2.98 (0.00291)	3.00(0.00269)	3.02(0.00250)	3.04(0.00235)	3.06(0.00222)	3.07(0.00212)
0.010	3.05(0.00230)	3.07(0.00212)	3.09(0.00197)	3.11(0.00185)	3.13(0.00175)	3.14(0.00166)

Table 2

Deshpande et al. (2010): Relative Risk (RR) estimation for Necrotizing Enterocolitis in Preterm Babies. Looks after 5,9,12, and 15 trials. $\Omega=\{5,9,12,15\}$
 $\gamma=0.0184$ (Maximum looks=4)

Study #	F/N Control	F/N Probiotics	Cum Est Log(RR) (SE)	$ T_K $ (P-value: non-sequential: if $< \gamma$, stop)	Cum Point est. of RR	Conf Interval
1	0/46	0/45				
2	8/290	4/295				
3	6/36	5/51				
4	10/73	1/72				
5	10/187	2/180	-1.13(0.34)	3.35(0.044)	0.32	0.066-1.59
6	3/41	1/39				
7	1/17	2/21				
8	3/31	0/38				
9	14/217	4/217	-1.14(0.23)	4.99(0.0016)	0.32	0.16-0.65
10	15/95	5/91				
11	1/49	2/45				

F=Failed N=Sample Size

Table 3

Performance of Pocock method: Probability (in percent) that at all looks, confidence intervals contain their true means. Ideal value is 95.0. Entries are Percentiles in 10,000 scenarios, each simulated 100,000 times. Increments refer to amount of information added per look.

Max Looks	Percentile	Random Increments	Increasing Increments	Decreasing Increments
2	Min	93.9	93.9	94.7
	10 th	94.2	94.3	94.9
	50 th	94.9	94.6	95.0
	90 th	96.0	94.9	95.3
	Max	96.7	95.1	96.4
3	Min	93.5	93.5	94.7
	10 th	94.2	93.9	95.0
	50 th	95.0	94.2	95.2
	90 th	96.0	94.6	95.7
	Max	97.2	95.0	96.9
4	Min	93.3	93.5	94.7
	10 th	94.2	93.9	95.1
	50 th	95.1	94.2	95.4
	90 th	95.9	94.6	95.9
	Max	97.5	95.0	97.1
5	Min	93.3	93.2	94.9
	10 th	94.3	93.6	95.4
	50 th	95.2	93.9	95.8
	90 th	96.0	94.4	96.3
	Max	97.7	95.0	97.3
6	Min	93.2	93.0	94.9
	10 th	94.3	93.5	95.6
	50 th	95.2	93.8	96.1
	90 th	96.0	94.3	96.5
	Max	97.5	95.0	97.4

Hypothetical Population Parameters for Inverse Weighted Sequential Meta-analysis. **Between Study Variance after 3 Studies=0**, after 4 studies=3

Table 4

Study Order	Within Study Mean	Within Study Variance	Variance 3-Studies	Weights 3 Studies	Variance 4-Studies	Weights 4 Studies (Relative Wt)
1	2	3	3+0=3	0.182	3+3=6	0.192 (0.270)
2	2	2	2+0=2	0.273	2+3=5	0.231 (0.325)
3	2	1	1+0=1	0.545	1+3=4	0.288 (0.405)
4	-2	1			1+3=4	0.288

Variance of study is within study variance+between study variance. Weights are inversely proportional to variance of study. Variance of estimate is the sum of the square of the weight times the variance for the study. In the last column relative weights of the first three studies (made to add to 1.00) are in () and differ substantially from the weights used after three studies.

Variance of Estimate: After 3 studies (4 studies) 0.545 (1.15). Perceived information went backward. Increment of information from study 3 to 4 is not independent of the information accrued through study 3.

Table 5

Hypothetical example showing independent increments

Study	N	Mean	N*Mean	Study	N	Mean	N*Mean
1	100	3	300	7	200	2	400
2	80	2	160	8	150	1	150
3	50	0	0	9	90	3	270
4	110	3	330				
5	70	2	140				
6	50	2	100				
Total	460		1030	Total	440		820

Estimate after six studies= $1030/460=2.239$ [These are calculated per equation (5)]

Estimate for studies 7–9: $820/440=1.864$

Estimate after 9 studies: $1850/900=2.056$ (ignoring increments)Sample size weighted estimate by increment: $[460(1030/460)+440(820/440)]/900=1850/900=2.056$ (Unchanged from ignoring increments)