Adaptive and Repeated Cumulative Meta-Analyses for Safety Signal Detection during a New Drug Development Process

Hui Quan, Yingqiu Ma, Yan Zheng, Meehyung Cho, Christelle Lorenzato, Carole Hecquet
Biostatistics and Programming, Sanofi, 55 Corporate Drive Bridgewater, NJ 08807

Abstract

During a new drug development process, it is desirable to use cumulative data from all completed studies sequentially to timely detect potential safety signals. For this purpose, repeated meta-analyses can be performed on combined data from multiple completed studies. Moreover, if data from the originally planned program are not enough for ensuring power to test a specific hypothesis, adaptation in total sample size to increase the amount of safety data by adding new studies into the program can also be conducted. Without appropriate adjustment, Type I error rate will be inflated because of the repeated analyses and sample size adaptation. In this paper, we provide a systematic discussion on potential issues associated with adaptive and repeated cumulative meta-analyses conducted during a drug development process. We consider both frequentist and Bayesian approaches. Simulation results are provided to compare the performances of different methods. A new drug development example is used to demonstrate the application of the methods.

Keywords: conditional power, combination test, Type I error rate control, fixed and random effect models, adaptive Bayesian design, non-inferiority.

1. Introduction

In a new drug development program, there are always multiple randomized controlled clinical trials. Most individual clinical trials are designed with primary objectives to evaluate treatment effects on specific primary endpoints. With very sparse safety data from individual studies, these trials are generally not large enough for detecting potential safety signals. The sponsor usually performs a meta-analysis using aggregated data from all studies in the drug development program for an integrated summary of safety at the end of the program. If substantial safety concern is raised during the review of the integrated summary of safety, the sponsor may give up and not file the new drug application to any health authorities. Thus, all resources put forward to the new drug development program will be wasted.

For example, a compound was recently developed for treating chronic insomnia characterized by difficulty with sleep maintenance. The Phase I clinical program included 34 studies with 867 subjects. The Phase II clinical program included 7 clinical trials with 1486 patients and the Phase III program consisted of 3 completed studies with 2715 patients and two ongoing studies with 919 patients. Results of Phase II and III studies clearly demonstrated that the experimental drug improved sleep maintenance with significant effects on multiple efficacy endpoints. However, as a part of a review of the safety data from Phase II-III completed and ongoing trials for integrated summary of safety, it was observed that 27 patients treated with the experimental drug developed diverticulitis versus none in patients treated with placebo. The most frequently associated symptoms with diverticulitis were abdominal pain and changes in bowel motility especially constipation. The excess of diverticulitis occurrences in treated patients was identified as a safety signal. After several rounds of reviews with input from external

experts including detailed benefit-risk assessment, the sponsor decided not to file a new drug application with results from the program. Consequently, several years' effort and a lot of resource spent on the development program were wasted. The questions were then whether the sponsor should do things differently and what should be done. The intuitive thinking was to perform repeated meta-analyses on cumulative safety data during the drug development process to timely detect potential safety signal and plan the development strategies accordingly.

There is extensive discussion on meta-analysis in literature [1-2]. Meta analysis is usually performed only once after all studies' completions and data from all studies becoming available. Then there will be no issue of Type I error rate inflation and no need to adjust the inference statistic for Type I error rate control. If however, repeated meta analyses are conducted sequentially on cumulative data during a new drug development process and adaptation in total sample size is also performed in order to have required amount of safety data for a specific objective, adjustment is necessary if we want to appropriately control the Type I error rate. There is very limited discussion on adaptive and repeated meta-analyses in literature [3-9]. The objective of this research is to conduct a systematic review on potential issues and methodologies associated with adaptive and repeated cumulative meta-analyses during a new drug development process.

The paper is organized as follows. In Section 2, we discuss data source for repeated cumulative meta-analyses. Based on the Clinical Development Plan, the number of cumulative meta-analyses can be planned and specified. Section 3 is used to outline the hypotheses, methods for hypothesis testing, estimation of treatment effect, conditional power calculation, power comparison as well as the analysis of binary and time-to-event endpoints. The Bayesian approach will be focused in Section 4. A new drug development example is used in Section 5 to illustrate the application of repeated cumulative meta-analyses for detecting adverse effect on diverticulitis. The paper is concluded with remarks in Section 6.

2. Data

Before the start of a new drug development, the clinical team always first has a Clinical Development Plan (CDP). The plan outlines the targeted patient population, the targeted indication(s) for the experimental drug, the number of studies, the primary objectives and endpoints for individual studies, the corresponding sample sizes for individual studies, and timelines for the completions of individual studies and so on. Data are cumulated over time as study is completed one after another. These cumulative data can be used in cumulative meta-safety analyses. If any substantial safety signal is detected, the sponsor can make timely decision on whether or not to stop the drug development program and save resource. In general, interim data of a study will not be unblinded and used in the meta-analyses. Interim safety data of a study should be monitored by study Data Monitoring Committee (DMC) to maintain the integrity of the study. Nevertheless, the sponsor should provide safety results of other completed studies to the DMC so that the DMC has all necessary information to make wise decision.

With information from the CDP, the clinical team can plan repeated meta-analyses based on the amount of information available sequentially overtime. Some studies such as small scaled Phase I studies on healthy subjects may not be included in the meta-analyses. To timely detect potential safety signal, a cumulative meta-analysis should be performed whenever there are substantial additional safety data, e.g., when a group of relatively

small studies are completed or a major Phase III study is completed. A stratified analysis approach is usually used for meta-analysis. Small studies in a chronological order can be combined to form a stratum and each Phase III study can form a stratum.

Even though data from individual patients are available to the sponsor, for most meta-analysis methods, only the summary statistics from individual strata are needed. Suppose δ is the parameter for measuring the overall treatment effect. For a safety endpoint, it is assumed that the smaller the δ the better the treatment outcome. Based on the CDP, let K be the planned total number of meta-analyses during the development process, further $\hat{\delta}_i$ and $\hat{\sigma}_i^2$ be the estimates of δ_i and the corresponding variances σ_i^2, respectively, where δ_i is the counter part of δ for the ith stratum or meta-analysis, i=1, …, K. Then, asymptotically, $\hat{\delta}_i \sim N(\delta_i, \sigma_i^2)$. For a fixed effect model, δ_i's are fixed parameters. For a random effect model, δ_i's are random variables and assumed to follow $\delta_i \sim N(\delta, \tau^2)$, where τ^2 is the between strata variance. If $\delta_1 = \delta_2 = ... = \delta_K = \delta$ or τ^2 =0, treatment effects across the strata are homogeneous or consistent.

3. Hypothesis testing and estimation

3.1. Hypotheses

Suppose a formal hypothesis testing approach is used for safety signal detection on a specific endpoint. If multiple endpoints are considered simultaneously instead, inference on individual endpoints is performed first, then multiplicity adjustment will be applied to the p-values adjusted for adaptive and repeated cumulative meta-analyses. There are two types of hypothesis testing settings. One is to test whether there is any treatment effect compared to a control using the so called superiority hypothesis. Another is to test whether the treatment effect is not greater than a pre-specified margin using the so called non-inferiority hypothesis. For the first hypothesis testing setting, the null and alternative hypotheses are

$$H_0 : \delta = 0 \qquad \text{versus} \qquad H_a : \delta > 0, \qquad (1)$$

where as mentioned earlier δ is the parameter for measuring the overall treatment effect. For a fixed effect model, δ can be treated as a weighted average of δ_i's. If H_0 is rejected, there is significant treatment effect. However, even there truly is treatment effect at certain degree, H_0 may not be rejected because of the small power for safety analysis. Care is needed to interpret the results in such a scenario. This setting is usually used when there is no prior knowledge regarding the potential treatment effect on the endpoint.

With the second hypothesis testing setting, the non-inferiority null and alternative hypotheses are

$$H_0^{'} : \delta \geq \Delta \qquad \text{versus} \qquad H_a^{'} : \delta < \Delta, \qquad (2)$$

where Δ>0 is the pre-specified margin for the evaluation. For example, FDA issued a guideline for new diabetic drug development [10]. All sponsors have to demonstrate the safety of the compounds on cardiovascular (CV) events based on non-inferiority assessment of not increasing the CV risk by more than 80% for the preliminary drug approval and not increasing the CV risk by more than 30% for the final drug approval. We do not expect an experimental drug to have a safety profile of better than the control particularly the placebo control. Hence, a Δ=0 in (2) is not considered at least at the

planning stage. If, nevertheless, $\Delta=0$ in (2) is considered, the endpoint should be treated as an efficacy endpoint rather than a safety endpoint. In general, for efficacy endpoint, meta-analysis is used to get a more precise estimate of the treatment effect and is not used for making efficacy claim. Efficacy claim should be based on data from well designed individual study. However, meta-analysis on efficacy endpoint can be used to generate hypothesis for future study. For instance, result from meta-analysis on mortality can be used to generate hypothesis for a future outcomes trial.

3.2. Methods for testing the hypotheses

When repeated meta-analyses are performed based on cumulative data, test of (1) will have an inflated Type I error rate if no alpha adjustment is performed. This Type I error rate increases with the number of meta-analyses. With a large Type I error rate, we could mistakenly claim the adverse effect for the drug or even kill a very valuable drug with a large chance. On the other hand, there is some legal implication if the sponsor continues the new drug development program even the treatment effect on the safety endpoint has already be demonstrated to be 'significant' without careful interpretation. Patients in the future potentially could sue the company. Nevertheless, if the Type I error rate is controlled under a very stringent criterion; a toxic drug may be claimed as a safe drug because of the mere failure of rejecting H_0 and be put into the market. Therefore, we have to carefully balance both the protection of patients from harmful drugs and the provision of new safe and effective medicines to patients. Sometimes, even when the treatment has significant effect on the safety endpoint, the drug may still be a viable treatment option based on benefit risk assessment and be put onto the market. The rationale for important decision making including critical e-mail exchanges should be documented in order to avoid legal complication in the future.

Several authors have discussed methods for Type I error rate control for cumulated meta-analyses. Lan et al. [3] and Hu et al. [4] propose to use the law of iterated logarithm to 'penalize' the Z value of the test statistic to account for multiple tests. The γ used in the penalization is obtained through simulation. It depends on the method for the analysis (e.g., risk difference, relative risk and odds ratio for binary endpoint) and the nominal significance level. A larger γ will result in a smaller Type I error rate. The number of meta-analyses can also impact the Type I error rate. These authors do not pre-specify the number of meta-analyses. Actually, they do not limit the number of meta-analyses since this number is usually subject to change when the sponsor adds new studies into the development program. Jennison and Turnbull [5] apply combination methods to meta-analysis. Here, we focus on how to apply the combination methods in a sequential way to cumulative meta-analyses.

Suppose K is the fixed total number of meta-analyses (or strata or the number of groups of studies) specified at the planning stage and P_i is the p-value of the ith stratum. Based on Fisher's p-value combination method [11],

$$T_k = -2\log(P_1P_2...P_k) \sim \chi^2_{2k} \quad \text{under } H_0$$

where $1 \le k \le K$, χ^2_{2k} denotes a chi-square distribution with $2k$ degrees of freedom since P_i's are independent and follow U(0,1) so that all $-\log(P_i) \sim \text{Exp}(1) \sim \frac{1}{2}\chi^2_2$ under H_0, i=1, …, k. For a pre-specified non-decreasing alpha spending function $\alpha(k)$ such that

$\alpha(K) = \alpha$, treatment effect is significant with Type I error rate appropriately control if the test statistic at the *k*th meta analysis exceeds critical value c_k where

$$\Pr(T_1 \geq c_1 \mid H_0) = \alpha(1)$$
$$\Pr(T_1 < c_1, T_2 \geq c_2 \mid H_0) = \alpha(2) - \alpha(1)$$

and

$$\Pr(T_1 < c_1, ..., T_{k-1} < c_{k-1}, T_k \geq c_k \mid H_0) = \alpha(k) - \alpha(k-1). \quad (3)$$

Even though theoretically possible, it is not very straight forward to calculate the above joint probabilities because of the correlations among these test statistics with chi-square distributions.

During the drug development process, the original CDP may be modified based on the need as some studies are added into the program and some studies are cancelled from the program. Depending on the reason for changing the drug development plan, some adjustment in the analysis procedure may be necessary to control the Type I error rate. After the *k*th analysis, suppose we add or delete some studies. The reason for adding or deleting these studies could be totally independent with the observed safety results of the completed studies (clearly documented). For example, some new studies may be added for obtaining additional efficacy indications. Then, we can change the number of cumulative meta analyses from K to K'. The remaining unspent alpha after the *k*th analysis will be spent over the K'-*k* analyses, which will not inflate the Type I error rate. Unless the safety effect is already significant based on the previous analyses and the sponsor decides to terminate the program, there should be no reason to delete some studies based on the observed safety results from the previous studies. Nevertheless, additional studies could be added for obtaining more safety data based on the observed safety results of the previous *k* analyses (or P_i, *i*=1, …, *k*). In such a case, the pre-fixed K should not be changed for not inflating the Type I error rate. P-values for studies completed between the originally planned (k' -1)th and k' th strata should be appropriately combined with the p-value of the k'th stratum to obtain a new $P_{k'}$, k'=*k*+1, …K to form the cumulative meta-analysis test statistics. P-values for studies completed after the originally planned last stratum (*K*th stratum) should also be combined with the p-value of the *K*th stratum.

Let's see an example of the original K of 2. Suppose the final number of studies after the completion of the first study depends on the observed P_1. For instance, if $P_1 > 0.2$, only the originally planned second study will be conducted (*K*=2). However, if $P_1 \leq 0.2$, one study will be added besides the originally planned second study. Denote P_2 as the p-value from the originally planned second study and Q_2 as the p-values for the added study. If only the originally planned second study is conducted, Q_2 does not exist. The correct test statistic for the second cumulative meta-analysis should be $-2\log(P_1 P_{2'}) \sim \chi_4^2$ where $P_{2'}$ is P_2 if only the originally planned second study is conducted (or $P_1 > 0.2$) and $P_{2'}$ is the p-value derived from $\Pr(\chi_4^2 \geq -2\log(P_2 Q_2))$ if a total of three studies are conducted (or $P_1 \leq 0.2$) since $P_{2'}$ always follows U(0,1) under H_0 regardless of the value of P_1 [5].

Rather than combining p-values, another approach for repeated cumulative meta-analysis is to combine test statistics [12]. Denote $Z_i = \hat{\delta}_i / \hat{\sigma}_i$ the test statistic and w_i the pre-specified weight for Z_i , i=1, …, K; such that $\sum_{i=1}^{K} w_i^2 = 1$. Then, asymptotically $V_k = \sum_{i=1}^{k} w_i Z_i / \sqrt{\sum_{i=1}^{k} w_i^2} \sim N(0,1)$ under H_0, k=1, …, K and can be used for the cumulative meta-analyses. If additional studies are added during the development process, new studies completed between the planned (k'-1)th and k'th strata can be combined with the original k'th stratum to form a new $Z_{k'}$. For example, before we see any data from these new studies and data from the k'th stratum, a separate new set of weights can be specified. These weights are used to combine test statistics from these new studies and the k'th stratum to form $Z_{k'}$. The efficient weights should be proportional to the square root of the anticipated amount of information (e.g., sample size or the number of events) provided by the corresponding studies.

Besides detecting significance of treatment effect through hypothesis testing, we are often also interested in the estimation of the magnitude of treatment effect. The combination test statistic approach rather than the combined p-value approach may be more appropriate for this purpose. With confidence interval, the combination test statistic approach is particularly more flexible for testing the non-inferiority hypothesis (2). Suppose, asymptotically, $\hat{\delta}_i \sim N(\delta, \sigma_i^2)$ (a common δ_1 , δ_2 , …, δ_K) and $\hat{\sigma}_i^2$ is an estimate of σ_i^2. Then, asymptotically,

$$Z_i^{'} = (\hat{\delta}_i - \delta) / \hat{\sigma}_i \sim N(0,1) \text{ and } T_k^{'} = (\sum_{i=1}^{k} w_i Z_i^{'}) / \sqrt{\sum_{i=1}^{k} w_i^2} \sim N(0,1) .$$

The estimate of δ at the kth meta-analysis is

$$\hat{\delta}^{(k)} = (\sum_{i=1}^{k} w_i \hat{\delta}_i / \hat{\sigma}_i) / (\sum_{i=1}^{k} w_i / \hat{\sigma}_i) .$$

If there is possibility of rejecting $H_0^{'}$ early, to control the Type I error rate, we cannot directly use the nominal confidence upper bound for testing $H_0^{'}$. The adjusted confidence upper bound for δ should be used. For the kth cumulative meta-analysis, it should be

$$UB^{(k)} = (\sum_{i=1}^{k} w_i \hat{\delta}_i / \hat{\sigma}_i + c_k^{'} \sqrt{\sum_{i=1}^{k} w_i^2}) / (\sum_{i=1}^{k} w_i / \hat{\sigma}_i)$$

where $c_k^{'}$ is the critical value for the kth meta analysis derived based on the joint distribution of $T_k^{'}$ and the alpha spending function, k=1, …, K (see (3) for a similar case). If $UB^{(k)} \le \Delta$, $H_0^{'}$ can be rejected and $H_a^{'}$ can be accepted with Type I error rate controlled. In reality, even there may be already enough data at the kth cumulative meta-analysis to claim the non-inferiority of the drug with Type I error rate appropriately controlled, the sponsor will not stop the ongoing studies and will continue to initial other studies accordingly to the original CDP for the planned claims. Therefore, to reserve all significance level for the final analysis, we probably should set $\alpha(k)$=0, k=1,…, K-1 and $\alpha(K) = \alpha$ to increase power for the non-inferiority assessment with the corresponding critical values of $c_k^{'} = \infty$, k=1,…, K-1 and $c_K^{'} = z_{1-\alpha}$, the $1-\alpha$ percentile of the standard normal distribution. Even that is the case; we still need to closely monitor safety data during the trial. If we find potentially not enough data to make the non-inferiority claim

based on conditional power calculation or other methods, we may add new studies into the program. Then the combination method discussed above with the pre-specified weights could be applied to the final analysis.

3.3. Conditional power calculation

During the new drug development process, as we closely monitor the cumulative safety data, we may change our objective and evaluate the corresponding power. When there is no treatment effect, power for rejecting null hypothesis H_0 in (1) is the type I error rate – in the sense that there will be no power. Anyway, rejecting H_0 is not the goal of the sponsor for safety evaluation. Thus, conditional power calculation for ultimately rejecting H_0 is rarely performed. When we detect significant treatment effect based on testing (1) with Type I error rate appropriately controlled or observe certain degree of treatment effect and have concern on the safety of the drug, the ultimate goal of the safety evaluation from the sponsor's perspective may be to demonstrate the non-inferiority of the drug through testing (2). We then need to assess the conditional power for this objective based on the observed result to see whether additional safety data from additional studies are necessary. Shih et al. [13] and Wang et al. [14] also considered adaptive non-inferiority assessment with conditional power but for a single trial.

Suppose the conditional power calculation is performed after the *k*th analysis. As mentioned earlier, no alpha will be spent at interim analyses for testing (2) and all the alpha will be reserved for the final analysis. Given the observed fixed results from the previous *k* analyses, the conditional power for demonstrating the non-inferiority is (since $\sum_{i=1}^{K} w_i^2 = 1$)

$$\mathrm{CP}(\delta) = \Pr\left(Z \le \frac{\Delta\sum_{i=1}^{K} w_i/\hat{\sigma}_i - \sum_{i=1}^{k} w_i\hat{\delta}_i/\hat{\sigma}_i - z_{1-\alpha} - \delta\sum_{i=k+1}^{K} w_i/\hat{\sigma}_i}{\sqrt{\sum_{i=k+1}^{K} w_i^2}} \,\middle|\, \delta, \sum_{i=1}^{k} w_i\hat{\delta}_i/\hat{\sigma}_i\right),$$

where asymptotically

$$Z = \frac{\sum_{i=k+1}^{K} w_i\hat{\delta}_i/\hat{\sigma}_i - \delta\sum_{i=k+1}^{K} w_i/\hat{\sigma}_i}{\sqrt{\sum_{i=k+1}^{K} w_i^2}} \sim N(0,1).$$

To have $1-\beta'$ conditional power, approximately

$$(\Delta-\delta)\sum_{i=k+1}^{K} w_i/\hat{\sigma}_i = \sum_{i=1}^{k} w_i(\hat{\delta}_i-\Delta)/\hat{\sigma}_i + z_{1-\alpha} + z_{1-\beta'}\sqrt{\sum_{i=k+1}^{K} w_i^2} \quad (4)$$

should hold. That is, we need to have enough data from the remaining strata such that $\hat{\sigma}_i$, $i=k+1, \ldots, K$ are small enough for (4) to hold. For non-inferiority assessment, δ in (4) could be set to zero. One could also use the estimate based on the previous *k* analyses

$$\hat{\delta}^{(k)} = \left(\sum_{i=1}^{k} w_i\hat{\delta}_i/\hat{\sigma}_i\right)/\left(\sum_{i=1}^{k} w_i/\hat{\sigma}_i\right)$$

to replace the δ in (4). If this $\hat{\delta}^{(k)}$ is very close to Δ or even greater than Δ already, using $\hat{\delta}^{(k)}$ in (4) will make it almost impossible to have enough data from the follow up studies for the desired conditional power.

3.4. Power comparison

In this section, we use simulation to compare power between the iterated algorithm method of Lan et al. [3] and the weighted combination test in Section 3.2 for testing H_0 (1). Here, the endpoint is assumed to follow a normal distribution and two scenarios are considered. One scenario consists of a total of 10 studies with sample size per treatment group of 100 for each of the first 7 studies and sample size per treatment group of 250 for each of the last 3 studies. Another scenario consists of a total of 25 studies with sample size per treatment of 100 for each of the first 15 studies and sample size per treatment group of 250 for each of the last 10 studies. The pre-specified weights w_i 's are $\sqrt{2/29} = \sqrt{100/1450}$, i=1, …, 7 and $\sqrt{5/29} = \sqrt{250/1450}$, i=8, 9, 10 for the first scenario, and are $\sqrt{1/40}$, i=1, …, 15 and $\sqrt{1/16}$, i=16, …25 for the second scenario. For each scenario, two additional cases are considered. One is to add one study that is combined with the last study for the weighted combination test. Another is to add two studies: one study is combined with the 8th study for the first scenario and with the 21th study for the second scenario; and another study is combined with the last study for the weighted combination test. Thus, the pre-specified number of cumulative meta-analyses (strata) and weights are not changed. The alpha spending function of $\alpha(k) = 1 - \Phi(z_\alpha / \sqrt{k/K})$ (one-sided) is used for the repeated cumulative meta-analyses for the weighted combination test to control the overall Type I error rate. For some cases, the differences in power between the two approaches can be larger than 15% (Table 1).

Table 1a. Power comparison between the iterated algorithm method ($\gamma = 1$) and the weighted combination test

	10 studies		Adding 1 study		Adding 2 studies	
δ	Iterated	Weighted	Iterated	Weighted	Iterated	Weighted
0.00	0.024	0.025	0.025	0.025	0.026	0.025
0.02	0.054	0.071	0.059	0.074	0.063	0.079
0.04	0.117	0.168	0.133	0.181	0.148	0.200
0.06	0.232	0.328	0.268	0.355	0.305	0.397
0.08	0.398	0.526	0.459	0.571	0.518	0.629
0.10	0.591	0.724	0.668	0.771	0.735	0.824
0.12	0.770	0.869	0.839	0.904	0.889	0.937
0.14	0.895	0.950	0.939	0.969	0.965	0.983
0.16	0.961	0.985	0.982	0.992	0.992	0.997
0.18	0.989	0.997	0.996	0.999	0.999	1.000
0.20	0.998	1.000	0.999	1.000	1.000	1.000

Table 1b. Power comparison between the iterated algorithm method ($\gamma = 1.1$) and the weighted combination test

	25 studies		Adding 1 study		Adding 2 studies	
δ	Iterated	Weighted	Iterated	Weighted	Iterated	Weighted
0.00	0.023	0.025	0.023	0.025	0.023	0.025
0.02	0.076	0.128	0.079	0.129	0.083	0.133
0.04	0.235	0.386	0.249	0.394	0.263	0.410
0.06	0.526	0.717	0.555	0.729	0.583	0.750
0.08	0.816	0.925	0.841	0.933	0.864	0.944
0.10	0.959	0.989	0.970	0.991	0.978	0.994

3.5. Binary and time-to-event endpoints

Even for binary endpoint and time-to-event endpoint, as long as the method for deriving the individual p-values for individual strata is valid, the p-value combination method is always valid no matter whether the event rates and the numbers of events are small (perhaps through exact method) or large. Nevertheless, as Lan et al. [3] and Hu et al. [4] considered continuous endpoint and binary endpoint separately when they proposed the use of the law of iterated logarithm for cumulative meta-analyses, we need also to provide additional details when we use the combination statistic methods for binary endpoint and time-to-event endpoint. For a binary endpoint, suppose the 2x2 table for the *i*th stratum is

	Event	Non-Event	
Treatment	a_{1i}	b_{1i}	n_{1i}
Control	a_{0i}	b_{0i}	n_{0i}
	m_{1i}	m_{0i}	n_i

Let r_{1i} and r_{0i} be the underlying event rates for the treatment group and control group for the *i*th stratum, respectively. Then, the corresponding estimators are $\hat{r}_{1i} = a_{1i} / n_{1i}$ and $\hat{r}_{0i} = a_{0i} / n_{0i}$. The variances/asymptotic variances of the estimators of the risk difference $\hat{r}_{1i} - \hat{r}_{0i}$, log risk ratio $\log(\hat{r}_{1i} / \hat{r}_{0i})$ and log odds ratio $\log(\hat{r}_{1i}(1-\hat{r}_{0i})/(\hat{r}_{0i}(1-\hat{r}_{1i})))$ are $\frac{r_{1i}(1-r_{1i})}{n_{1i}} + \frac{r_{0i}(1-r_{0i})}{n_{0i}}$, $\frac{(1-r_{1i})}{n_{1i}r_{1i}} + \frac{(1-r_{0i})}{n_{0i}r_{0i}}$ and $\frac{1}{n_{1i}r_{1i}} + \frac{1}{n_{1i}(1-r_{1i})} + \frac{1}{n_{0i}r_{0i}} + \frac{1}{n_{0i}(1-r_{0i})}$, respectively. For a fixed effect model, we may assume $r_1 = r_{1i}$ and $r_0 = r_{0i}$ for all *i*'s for assessing the overall treatment effect to simplify all the computations. If all studies use balanced design ($n_{1i} = n_{0i}$ for all *i*'s), all these variances are $1/n_{1i}$ multiplied by constants. Therefore, the efficient pre-specified weights can be simply the square root of the anticipated fractions of the sample sizes of the strata (i.e., $w_i = \sqrt{n_{1i} / \sum_{j=1}^{K} n_{1j}}$) regardless of the event rates for the two treatment groups and whether risk difference, risk ratio or odds ratio is used for the analysis. Sometimes, we may use unbalanced designs for some studies ($n_{1i} = \eta_i n_{0i}$) in order to have more safety data for the experimental drug in some particular patient populations. Then it will be difficult to pre-specify the efficient weights in general unless for some specific scenarios. For example, if the event rates for the two treatment groups are anticipated to be close, the efficient weights could be $w_i = \sqrt{n_{1i}\eta_i/(1+\eta_i) / \sum_{j=1}^{K}(n_{1j}\eta_j/(1+\eta_j))}$. When the event rates and the numbers of events for individual strata are very small particularly if the numbers of events for some groups within some strata are zero, the asymptotic normal distributions for the combination statistic method may not be valid and we need alternative method for the analysis. For example, some type of continuity correction should be made. Also, it is a common practice to use a pre-specified rule to combine data from adjacent small strata in analysis.

When studies and patients have very different study/treatment durations, to take treatment exposure into account, we may use method for time-to-event endpoint in data analysis. The non-parametric log rank test is often used for deriving p-value while the semi-parametric proportional hazards model is often used for estimating the magnitude of treatment effect. For rare events, the exposure adjusted event rate method is also used in data analysis. The assumption of constant hazard rate for this method is reasonable for rare and chronic events, e.g., cardiovascular events. For illustration purpose, particularly for pre-specifying the weights for the weighted combination test, let's focus on the exposure adjusted event rate approach. The method for defining the weights in the following can also be applied to log rank test or the proportional hazards model approach.

Suppose λ_1 and λ_0 are the overall hazard rates of the treatment and control groups, respectively. Let $\hat{E}_{1i}$ and $\hat{E}_{0i}$ be the observed values of the expected numbers of events E_{1i} and E_{0i}, U_{1i} and U_{0i} be the observed total exposures for the treatment group and control groups for the *i*th stratum, respectively. Then λ_1 and λ_0 can be estimated by $\hat{\lambda}_{1i} = \hat{E}_{1i} / U_{1i}$ and $\hat{\lambda}_{0i} = \hat{E}_{0i} / U_{0i}$. Asymptotically, the estimated log hazard ratio

$$\hat{\delta}_i' = \log(\hat{\lambda}_{1i} / \hat{\lambda}_{0i}) \sim N(\delta, 1/E_{1i} + 1/E_{0i}) \text{ . (5)}$$

If E_{1i} and E_{0i} are similar (particularly for balanced design), (5) becomes

$$\hat{\delta}_i' = \log(\hat{\lambda}_{1i} / \hat{\lambda}_{0i}) \sim N(\delta, 4/E_i) \text{ (6)}$$

where E_i is the total expected number of events for the two treatment groups combined for the *i*th stratum. We then use the pre-specified weights $w_i = \sqrt{(E_{1i}E_{0i})/(E_{1i}+E_{0i}) / \sum_{j=1}^{K}(E_{1j}E_{0j})/(E_{1j}+E_{0j})}$ or $w_i = \sqrt{E_i / \sum_{j=1}^{K} E_j}$ in the analysis if all studies use balanced design. The estimator of the asymptotic variance is $\hat{\sigma}_i'^2 = 1/\hat{E}_{1i} + 1/\hat{E}_{0i}$ or $\hat{\sigma}_i'^2 = 4/\hat{E}_i$. With these quantities, we can use the formulas to perform the repeated meta-analyses. Similar to the case of a binary endpoint, if the event rates and the numbers of events for some treatment groups in some strata are very small, (5) and (6) may not hold. We need to combine data from multiple small adjacent strata using a pre-specified rule in the analysis.

4. Bayesian approach

Many authors have discussed the use of Bayesian methods for fixed meta-analysis [15-18]. Bayesian methods are also very natural and useful tools for synthesizing cumulative information in a sequential way. In this section, we will discuss their use in a repeated cumulative meta-analysis setting.

For a one level Bayesian model (as compared to the hierarchical Bayesian model), suppose $f(\theta)$ is the prior distribution for θ. With data x_1 for the first meta-analysis, we can obtain the posterior distribution for θ as $f(\theta \mid x_1) \propto f(x_1 \mid \theta) f(\theta)$. Then, with data x_2 from the second stratum, we can use $f(\theta \mid x_1)$ as the updated priori distribution to obtain an updated posterior distribution for the second meta-analysis as $f(\theta \mid x_1, x_2) \propto f(x_2 \mid \theta) f(\theta \mid x_1) \propto f(x_1 \mid \theta)\ f(x_2 \mid \theta) f(\theta)$. Clearly, it is exactly the same posterior distribution based on the original $f(\theta)$ as the prior distribution and all available

data of x_1 and x_2 from the two strata**.** In general, after we have data for *k* strata, the posterior distribution is

$$f(\theta \mid x_1, x_2, \ldots, x_k) \propto f(x_k \mid \theta) f(\theta \mid x_{1,} \cdots, x_{k-1}) \propto \prod_{i=1}^{k} f(x_i \mid \theta) f(\theta)$$

where the likelihood part $\prod_{i=1}^{k} f(x_i \mid \theta)$ is the same no matter whether the analysis is based on the updated prior distribution and the newly available data or the original prior distribution and all available data. Therefore, the easiest way is to simply perform the Bayesian analysis using the original prior distribution and all available data for each cumulative meta-analysis.

Let's see the case when the endpoint follows a normal distribution. Suppose $\delta \sim N(\mu_0, \sigma_0^2)$. When $\sigma_0^2 \to \infty$, it is a non-informative prior. For endpoint with a normal distribution and variance σ^2, the posterior distribution for the first meta-analysis is

$$(\delta \mid \mu_0, \hat{\delta}_1) \sim N(\frac{n_1 \hat{\delta}_1 \sigma_0^2 + \mu_0 2\sigma^2}{n_1 \sigma_0^2 + 2\sigma^2}, \frac{2\sigma^2 \sigma_0^2}{n_1 \sigma_0^2 + 2\sigma^2})$$

where n_i is the sample size per treatment group for the *i*th stratum. If $\sigma_0^2 \to \infty$, $(\delta \mid \mu_0, \hat{\delta}_1) \sim N(\hat{\delta}_1, \frac{2\sigma^2}{n_1})$. With data from the second stratum (fixed n_2), the updated posterior distribution is

$$(\delta \mid \mu_0, \hat{\delta}_1, \hat{\delta}_2) \sim N(\frac{(n_1 \hat{\delta}_1 + n_2 \hat{\delta}_2)\sigma_0^2 + \mu_0 2\sigma^2}{(n_1 + n_2)\sigma_0^2 + 2\sigma^2}, \frac{2\sigma^2 \sigma_0^2}{(n_1 + n_2)\sigma_0^2 + 2\sigma^2}) \quad (7)$$

As for the case of frequentist analysis, if sample size n_2 is a function of $\hat{\delta}_1$, (7) will not hold since $(\frac{n_1 \hat{\delta}_1 + n_2(\hat{\delta}_1)\hat{\delta}_2}{n_1 + n_2(\hat{\delta}_1)} \mid \delta)$ will no longer follow a distribution $N(\delta, \frac{2\sigma^2}{n_1 + n_2})$ as discussed in Section 3. We have to be careful on the method for deriving the posterior distribution.

Many authors use a hierarchical Bayesian model for fixed stratified and meta analysis of a binary endpoint. Using the notations in Section 3, the following are assumed under such a hierarchical model for a binary endpoint.

$$a_{1i} \sim Bin(n_{1i}, r_{1i}), \quad a_{0i} \sim Bin(n_{0i}, r_{0i})$$

$$\mu_i = \log it(r_{0i}), \quad \mu_i + \delta_i^* = \log it(r_{1i}), \quad \mu_i \sim N(\mu_0, \sigma_0^2), \quad \delta_i^* \sim N(\delta^*, \sigma_\delta^2), \; i=1, \ldots, K;$$

prior distributions $\mu_0 \sim N(0, c^2)$, $\delta^* \sim N(0, d^2)$, $\sigma_0^2 \sim$ InverseGamma(g_1 , g_2) and $\sigma_\delta^2 \sim$ InverseGamma(g_3 , g_4), where c^2, d^2 and g 's are all known constants [19]. Other prior distributions could also be applied. For example, some uses $\sigma_\delta^2 \sim$ Unif(0, s^2) [16]. The objective is to derive the updated posterior distribution of δ^* for measuring the magnitude of treatment effect at each meta-analysis. With this hierarchical model, the posterior distribution has no close form formula, a simulation approach (MCMC) should be applied. After obtaining the posterior distribution, we can evaluate probability Pr($\delta^* > a$) (for some value $a \geq 0$) or Pr($\delta^* \leq \Delta$) to decide whether there is treatment effect or the treatment is non-inferior to the control on the endpoint.

5. Example

In this section, we use the new drug development example introduced in Section 1 to illustrate some key points for repeated cumulative meta-analyses. Table 2 presents the data from the program based on the chronological order of study completions. For illustration purpose, the 11 studies were assumed to be divided into 5 strata.

Because of zero number of events in the control group, the first reaction on the analysis of this data set was to use the exact method that is nevertheless known to be conservative. When no alpha was spent for interim analyses and all alpha was reserved for the final analysis (the interim critical values for correlated chi-square test statistics are difficult to derive anyway), the combined p-value approach with the exact method for deriving p-values for individual strata had an adjusted p-value of 0.029 failing to claim significant treatment effect at level of 0.025 (one-sided) even though the pooled unstratified analysis provided an extremely small p-value of 1.37×10^{-5}. The reason for this could be that the exact method provided very conservative p-value for each stratum and then extremely conservative combined p-value.

Table 2. Data of diverticulitis events from the new drug development program

			Control			Treatment		
Study	Phase	Duration (days)	N	PYs*	# of Events	N	PYs*	# of Events
1	I in pats	14	6	0.23	0	13	0.46	0
2	II	14	21	0.80	0	22	0.84	0
3	II	42	112	14.08	0	213	27.01	1
4	II	42	25	1.40	0	73	4.87	0
Stratum 1								
5	II	168	105	22.72	0	324	68.36	3
6	II	28	19	1.29	0	18	1.19	0
7	II	28	119	9.12	0	231	17.44	0
8	II	8	70	8.70	0	134	17.39	0
Stratum 2								
9	III	42	311	34.68	0	297	32.81	1
Stratum 3								
10	III	84	295	60.79	0	850	176.15	14
Stratum 4								
11	III	84	345	67.20	0	617	123.43	8
Stratum 5								
12	III	28	143	10.52	0	140	10.08	0
13	III	42	315	35.69	0	321	36.15	0

* Patient-years

We considered the weighted combination test based on asymptotic normal distribution. With alpha spending function of $\alpha(k)=1-\Phi(z_\alpha/\sqrt{k/K})$, significance of treatment effect was detected at the 3rd cumulative meta analysis when significance level $\alpha=0.05$ and at the 4th cumulative meta analysis when significance level $\alpha=0.025$. Another way was to use the adding 0.2 events to each group and each stratum continuity correction approach (a total of 1 event for each treatment group for the 5 strata). With this approach, significant treatment effect was detected at the 4th cumulative meta-analysis at both significance levels

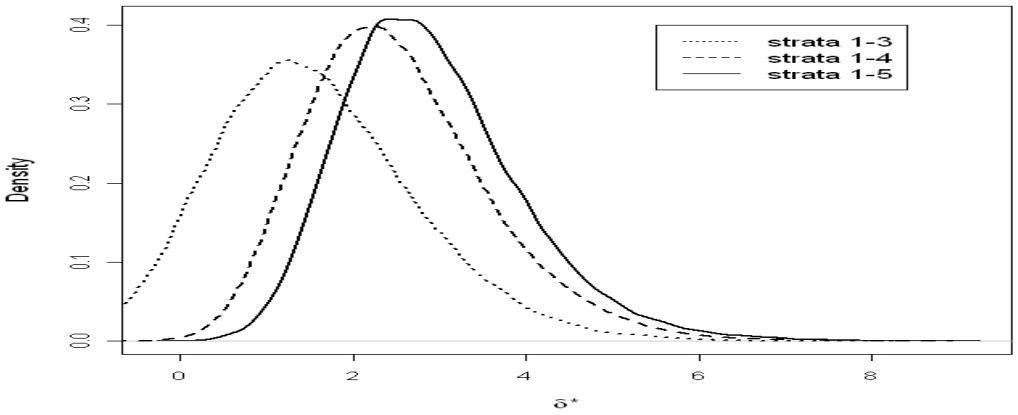
$\alpha = 0.025$ and 0.05. If risk ratio or odds ratio rather than difference of rates with adding 0.2 events continuity correction approach was applied, significant treatment effect was detected at the 5th cumulative meta-analysis at both significance levels $\alpha = 0.025$ and 0.05. The last two studies were added into the program after observing the results from the previous studies, which could be considered as a sample size adaptation. If these studies were combined with the original Stratum 5 (Study 11) in a stratified analysis to obtain test statistic $Z_{5'}$ that was used in the weighted combination test with the original weight w_5 assigned to Stratum 5, significant treatment effect was detected at the 5th cumulative meta analysis at significance level $\alpha = 0.05$ based on log risk or odds ratio.

The hierarchical Bayesian model in Section 4 was also used to analyze the data set with prior distributions $\mu_0 \sim N(0,10)$, $\delta^* \sim N(0,10)$, $\sigma_0^2 \sim$ InverseGamma(3,1) and $\sigma_\delta^2 \sim$ InverseGamma(3,1) [19]. Results are summarized in Table 3 and Figure 1 for cumulative meta-analyses of strata 1-3, 1-4 and 1-5. After the 4th cumulative meta-analyses, the 2.5 percentile of the posterior distribution for the log odds ratio exceeded 0 indicating a significant treatment effect. From Figure 1, there was also a clear shift from left to right from the third to fifth meta-analysis on the posterior density for the log odds ratio.

Table 3. Results from hierarchical Bayesian model for the log odds ratio

Strata	mean	SD	2.5%	median	97.5%
1-3	1.563	1.177	-0.4688	1.468	4.145
1-4	2.522	1.071	0.7727	2.404	4.947
1-5	2.903	1.041	1.219	2.784	5.282

Figure 1. Posterior density for the log odds ratio on diverticulitis events



Through this data set, it showed that when the numbers of events for individual treatments and strata were small, power for detecting treatment effect diminished based on a stratified analysis compared to a pooled analysis. For example, p-value of 1.37×10^{-5} via the pooled analysis could quickly increase to around 0.029 of a stratified analysis. Therefore, care should be exercised when determining the analysis strategy or multiple sensitivity analyses should be performed to confirm the results. Once treatment effect was detected, the sponsor could make adjustment of modifying the development strategy or re-arranging the priority of resources.

6. Discussion

It is unethical for a sponsor not to closely monitor cumulative safety data during a new drug development process in case the drug truly has substantial adverse effect. In addition, the sponsor surely wants to timely detect safety signals and terminate the development of an unsafe drug in order to save valuable resources. For this purpose, repeated and potentially adaptive cumulative meta-analyses should be planned as early as at the Clinical Development Plan (CDP) stage. Safety data from individual studies are usually sparse. Combining them in cumulative meta-analyses whenever substantial amount of new data becoming available during the drug development process will increase power for detecting safety signals. Nevertheless, different from the fixed meta-analysis that has been well studied, adaptive and repeated safety cumulative meta-analyses have some issues that worth additional attention.

Strong Type I error rate control associates with reduction in power. For safety analysis, balance the ability to detect safety signals and the inflation of Type I error rate is critical. From patient and health authority perspective, controlling Type I error rate for testing the non-inferiority hypothesis is more important than testing the superiority hypothesis in safety evaluation. If controlling Type I error rate is really desirable, the combination method is a convenient method even for adaptive and repeated cumulative meta-analyses. With this method, studies can be added to increase sample size even based on results of previous studies and the Type I error rate is still controlled.

With significant adverse effect observed, there could be multiple options: 1). The sponsor could stop all ongoing studies and terminate the whole drug development program immediately if the safety endpoint is a very important endpoint (e.g., mortality); 2). Nevertheless, drugs with certain side effects may still be valuable medicines. For less serious safety endpoint (e.g., gastrointestinal bleeding and depression), the sponsor could stop the initiations of new studies but continue those ongoing studies until their completions to get additional safety data for additional meta-analyses; 3). The sponsor could also continue all studies including those have not yet been started according to the original plan; 4). The sponsor could even design a stand alone safety study with pre-specified safety hypothesis (e.g., the non-inferiority hypothesis) and enough power to definitely address/confirm the safety concern or quantify the magnitude of the effect for benefit/risk assessment. The rationale for these actions should be carefully documented. No matter what follow-up option is applied, since significance of treatment effect has already been demonstrated with Type I error rate well controlled, potential additional follow up analysis will not inflate the Type I error rate. Type I error rate control may not be the ultimate goal during repeated cumulative meta-analyses on safety endpoint. However, knowing the precise error rate of a claim will help us to make wise decision.

References

1. Der Simonian R and Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trisls* 1986; 7: 177-188.
2. Sutton AJ, and Higgins JPT. Recent developments in meta-analysis. *Statistics in Medicine* 2008; 27: 625-650.
3. Lan KKG, Hu M and Cappelleri JC. Applying the law of iterated logarithm to cumulative meta-analysis of a continuous endpoint. *Statistica Sinica* 2003; 13: 1135-1145.

4. Hu M, Cappelleri JC and Lan KKG. Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clinical Trials* 2007; 4: 329-340.
5. Jennison C and Turnbull BW. Meta-analyses and adaptive group sequential designs in the clinical development process. *Journal of Biopharmaceutical Statistics* 2005; 15: 537-558.
6. Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F and Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *The New England Journal of Medicine* 1992; 327: 248-254.
7. Henderson WG, Moritz T, Goldman S, Copeland J and Sethi G. Use of cumulative meta-analysis in the design, monitoring, and final analysis of a clinical trial: a case Study. *Controlled Clinical Trials* 1995; 16: 331-341.
8. Berkey CS, Mosteller F, Lau J and Antman EM. Uncertainty of the time of first significance in random effects cumulative meta-analysis. *Controlled Clinical Trials* 1996; 17: 357-371.
9. Pogue JM and Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clinical Trials* 1997; 18: 580-593.
10. FDA, Guidance for Industry: diabetes mellitus-evaluating cardiovascular risk in new antidiabetic therapies to treat Type 2 diabetes, December 2008.
11. Fisher RA. *Statistical Methods for Research Workers* 4th e. London: Oliver and Boyd 1932.
12. Cui L, Hung HM, and Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; 55: 853-857.
13. Shih WJ, Quan H and Li G. Two-stage adaptive strategy for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* 2004; 23: 2781-2798.
14. Wang SJ, Hung HMJ, Tsong Y and Cui L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* 2001; 20: 1903-1912.
15. Smith TC, Spiegelhalter DJ and Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* 1995; 14: 2685-2699.
16. Warn DE, Thompson SG and Spiegelhalter DJ. Bayesian random effects meta-analysis of trials with binary outcomes: methods fort he absolute risk difference and relative risk scales. *Statistics in Medicine* 2002; 21: 1601-1623.
17. Leonard T and Duffy JC. A Bayesian fixed effects analysis of the Mantel-Haenszel model applied to meta-analysis. Statistics in Medicine 2002; 21: 2295-2312.
18. Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR and Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. *Statistics in Medicine* 2007; 26: 2479-2500.
19. Berry SM and Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics* 2004; 60: 418-426.