Chapter 17
Statistical Methods Used in Interim Monitoring

In Chap. 16, the administrative structure was discussed for conducting interim analysis of data quality and outcome data for benefit and potential harm to trial participants. Although statistical approaches for interim analyses may have design implications, we have delayed discussing any details until this chapter because they really focus on monitoring accumulating data. Even if, during the design of the trial, consideration was not given to sequential methods, they could still be used to assist in the data monitoring or the decision-making process. In this chapter, some statistical methods for sequential analysis will be reviewed that are currently available and used for monitoring accumulating data in a clinical trial. These methods help support the evaluation of interim data and whether they are so convincing that the trial should be terminated early for benefit, harm, or futility or whether it should be continued to its planned termination. No single statistical test or monitoring procedure ought to be used as a strict rule for decision-making, but rather as one piece of evidence to be integrated with the totality of evidence [1–6]. Therefore, it is difficult to make a single recommendation about which should be used. However, the following methods, when applied appropriately, can be useful guides in the decision-making process.

Classical sequential methods, a modification generally referred to as group sequential methods, and curtailed testing procedures are discussed below in some detail; other approaches are also briefly considered. Classical sequential methods are given more mathematical attention in several articles and texts which can be referred to for further detail [7–20].

Fundamental Point

Although many statistical techniques are available to assist in monitoring, none of them should be used as the sole basis in the decision to stop or continue the trial.

The original version of this chapter was revised. An erratum can be found at DOI 10.1007/978-3-319-18539-2_23

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-3-319-18539-2_17

Classical Sequential Methods

The aim of the classical sequential design is to minimize the number of participants that must be entered into a study. The decision to continue to enroll participants depends on results from those already entered. Most of these sequential methods assume that the response variable outcome is known in a short time relative to the duration of the trial. Therefore, for many trials involving acute illness, these methods are applicable. For studies involving chronic diseases, classical sequential methods have not been as useful. Detailed discussions of classical sequential methods are given, for example, by Armitage [20], Whitehead [18], and Wald [16].

The classical sequential analysis method as originally developed by Wald [16] and applied to the clinical trial by others such as Armitage [8, 9, 20] involves repeated testing of data in a single experiment. The method assumes that the only decision to be made is whether the trial should continue or be terminated because one of the groups is responding significantly better than the other. This classical sequential decision rule is called an "open plan" by Armitage [20] because there is no guarantee of when a decision to terminate will be reached. Strict adherence to the "open plan" would mean that the study could not have a fixed sample size. Very few clinical trials use the "open" or classical sequential design. The method also requires data to be paired, one observation from each group. In many instances, the pairing of participants is not appealing because the paired participants may be very different and may not be "well matched" in important prognostic variables. If stratification is attempted in order to obtain better matched pairs, each stratum with an odd number of participants would have one unpaired participant. Furthermore, the requirement to monitor the data after every pair may not be possible for many clinical trials. Silverman and colleagues [21] used an "open plan" in a trial of the effects of humidity on survival in infants with low birth weight. At the end of 36 months, 181 pairs of infants had been enrolled; 52 of the pairs had a discrepant outcome. Nine infants were excluded because they were un-matched and 16 pairs were excluded because of a mismatch. The study had to be terminated without a clear decision because it was no longer feasible to continue the trial. This study illustrates the difficulties inherent in the applying the classical sequential design for clinical trials.

Armitage [8] introduced the restricted or "closed" sequential design to assure that a maximum limit is imposed on the number of participants ($2N$) to be enrolled. As with the "open plan," the data must be paired using one observation from each study group. Criteria for early termination and rejection of no treatment effect are determined so that the design has specified levels of significance and power (α and $1-\beta$). This design was used in a comparison of two interventions in patients with ulcerative colitis [22]. In that trial, the criterion for no treatment effect was exceeded, demonstrating short-term clinical benefit of corticosteroids over sulphasalazine therapy. This closed design was also used in an acute leukemia trial, comparing 6-mercaptopurine with placebo (CALGB) [23]. This trial was

terminated early, with the statistic comparing remission rates crossing the sequential boundary for benefit after 21 pairs of patients.

Another solution to the repeated testing problem, called "repeated significance tests," was proposed by McPherson and Armitage [24] and also described by Armitage [20]. Although different theoretical assumptions are used, this approach has features similar to the restricted sequential model. That is, the observed data must be paired, and the maximum number of pairs to be considered can be fixed. Other modifications to the Armitage restricted plan [25–27] have also been proposed. This methodology plays an important role in a method to be described below, referred to as group sequential design.

The methods described above can in some circumstances be applied to interim analyses of censored survival data [25, 28–36]. If participants simultaneously enter a clinical trial and there is no loss to follow-up, information from interim analyses is said to be "progressively censored." Sequential methods for this situation have been developed using, for example, modified rank statistics. In fact, most participants are not entered into a trial simultaneously, but in a staggered fashion. That is, participants enter over a period of time after which events of interest occur, subject to an independent censoring process. The log-rank statistic, described in Chap. 15, may also be used to monitor in this situation.

The classical sequential approach has not been widely used, even in clinical trials where the time to the event is known almost immediately. One major reason is that for many clinical trials, if the data are monitored by a committee which has regularly scheduled meeting, it is neither feasible nor necessary for ethical reasons to perform an analysis after every pair of outcomes. In addition, classical sequential boundaries require an alternative hypothesis to be specified, a feature not demanded by conventional statistical tests for the rejection of the null hypothesis.

Group Sequential Methods

Because of limitations with classical sequential methods, other approaches to the repeated testing problem have been proposed. Ad hoc rules have been suggested that attempt to ensure a conservative interpretation of interim results. One such method is to use a critical value of 2.6 at each interim look as well as in the final analyses [1]. Another approach [37, 38] referred to as the Haybittle–Peto procedure, favors using a large critical value, such as $Z_i = +3.0$, for all interim tests ($i < K$). Then any adjustment needed for repeated testing at the final test ($i = K$) is negligible and the conventional critical value can be used. These methods are ad hoc in the sense that no precise Type I error level is guaranteed. They might, however, be viewed as precursors of the more formal procedures to be described below.

Pocock [39–41] modified the repeated testing methods of McPherson and Armitage [24] and developed a group sequential method for clinical trials which avoids many of the limitations of classical methods. He discusses two cases of special interest; one for comparing two proportions and another for comparing

mean levels of response. Pocock's method divides the participants into a series of K equal-sized groups with $2n$ participants in each, n assigned to intervention and n to control. K is the number of times the data will be monitored during the course of the trial. The total expected sample size is $2nK$. The test statistic used to compare control and intervention is computed as soon as data for the first group of $2n$ participants are available, and then recomputed when data from each successive group of $2n$ participants become known. Under the null hypothesis, the distribution of the test statistic, Z_i, is assumed to be approximately normal with zero mean and unit variance, where i indicates the number of groups ($i \leq K$) which have complete data. This statistic Z_i is compared to the stopping boundaries, $\pm ZN_K$ where ZN_K has been determined so that for up to K repeated tests, the overall (two sided) significance level for the trial will be α. For example, if $K = 5$ and $\alpha = 0.05$ (two-sided), $ZN_K = 2.413$. This critical value is larger than the critical value of 1.96 used in a single test of hypothesis with $\alpha = 0.05$. If the statistic Z_i falls outside the boundaries on the "i"-th repeated test, the trial should be terminated, rejecting the null hypothesis. If the statistic never falls outside the boundaries, the trial should be continued until $i = K$ (the maximum number of tests). When $i = K$, the trial would stop and the investigator would "accept" H_0.

O'Brien and Fleming [42] also discuss a group sequential procedure. Using the above notation, their stopping rule compares the statistic Z_i with $Z^* \sqrt{(K / i)}$ where Z^* is determined so as to achieve the desired significance level. For example, if $K = 5$ and $a = 0.05$, $Z^* = 2.04$. If $K \leq 5$, Z^* may be approximated by the usual critical values for the normal distribution. One attractive feature is that the critical value used at the last test ($i = K$) is approximately the same as that used if a single test were done.

In Fig. 17.1, boundaries for the three methods described are given for $K = 5$ and $\alpha = 0.05$ (two-sided). If for $i < 5$ the test statistic falls outside the boundaries, the trial is terminated and the null hypothesis rejected. Otherwise, the trial is continued until $i = 5$, at which time the null hypothesis is either rejected or "accepted". The three boundaries have different early stopping properties. The O'Brien–Fleming model is unlikely to lead to stopping in the early stages. Later on, however, this procedure leads to a greater chance of stopping prior to the end of the study than the other two. Both the Haybittle–Peto and the O'Brien–Fleming boundaries avoid the awkward situation of accepting the null hypothesis when the observed statistic at the end of the trial is much larger than the conventional critical value (i.e., 1.96 for a two-sided 5% significance level). If the observed statistic in Fig. 17.1 is 2.3 when $i = 5$, the result would not be significant using the Pocock boundary. The large critical values used at the first few analyses for the O'Brien–Fleming boundary can be adjusted to some less extreme values (e.g., 3.5) without noticeably changing the critical values used later on, including the final one.

Many monitoring committees wish to be somewhat conservative in their interpretation of early results because of the uncertainties discussed earlier and because a few additional events can alter the results substantially. Yet, most investigators would like to use conventional critical values in the final analyses, not requiring any penalty for interim analyses. This means that the critical value used in a

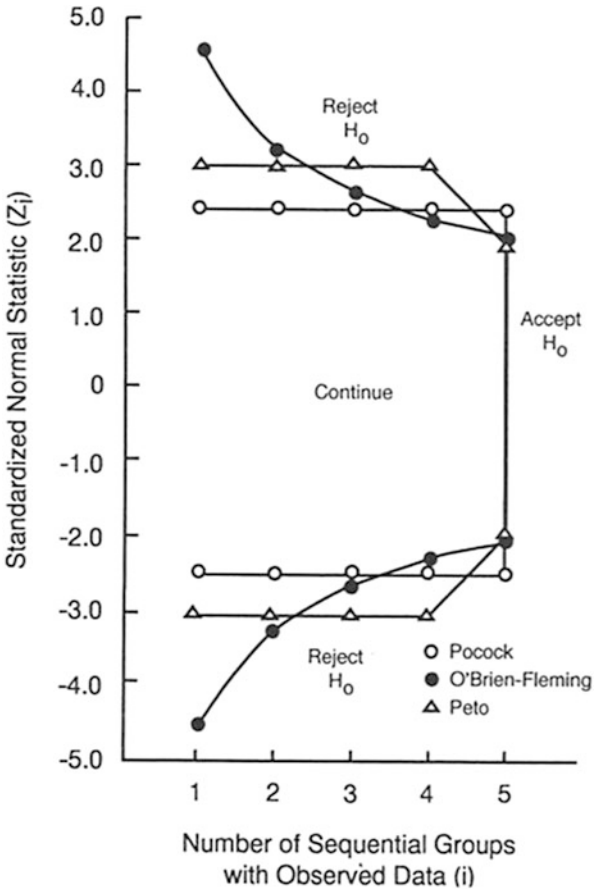

Fig. 17.1 Three group sequential stopping boundaries for the standardized normal statistic (Zi) for up to five sequential groups with two-sided significance level of 0.05 [64]

conventional fixed sample methods would be the same for that used in a sequential plan, resulting in no increase in sample size. With that in mind, the O'Brien–Fleming model has considerable appeal, perhaps with the adjusted or modified boundary as described. That is, the final critical value at the scheduled end of the trial is very close to the conventional critical value (e.g. 2.05 instead of 1.96) if the number of interim analyses is not excessive (e.g. larger than 10). The group sequential methods have an advantage over the classical methods in that the data do not have to be continuously tested and individual participants do not have to be "paired." This concept suits the data review activity of most large clinical trials where monitoring committees meet periodically. Furthermore, in many trials constant consideration of early stopping is unnecessary. Pocock [39–41] discusses the benefits of the group sequential approach in more detail and other authors describe variations [43–47].

In many trials, participants are entered over a period of time and followed for a relatively long period. Frequently, the primary outcome is time to some event.

Instead of adding participants between interim analyses, new events are added. As discussed in Chap. 15, survival analysis methods could be used to compare the experience of the intervention and the control arms. Given their general appeal, it would be desirable to use the group sequential methods in combination with survival analyses. It has been established for large studies that the log-rank or Mantel–Haenszel statistic [48–53] can be used. Furthermore, even for small studies, the log-rank procedure is still quite robust. The Gehan, or modified Wilcoxon test [54, 55], as defined in Chap. 15 does not always produce interim values with independent increments and so cannot be easily incorporated using the usual group sequential procedures. A generalization of the Wilcoxon procedure for survival data, though, is appropriate [56] and the survival methods of analyses can in general terms be applied in group sequential monitoring. Instead of looking at equal-sized participant groups, the group sequential methods described strictly require that interim analyses should be done after an additional equal number of events have been observed. Since monitoring committees usually meet at fixed calendar times, the condition of equal number of events might not be met exactly. However, the methods applied under these circumstances are approximately correct [57] if the increments are not too disparate. Other authors have also described the application of group sequential methods to survival data [58–61].

Interim log-rank tests in the Beta-Blocker Heart Attack Trial [62, 63] were evaluated using the O'Brien–Fleming group sequential procedure [42]. Seven meetings had been scheduled to review interim data. The trial was designed for a two-sided 5% significance level. These specifications produce the group sequential boundary shown in Fig. 17.2. In addition, the interim results of the log-rank statistic are also shown for the first six meetings. From the second analysis on, the conventional significance value of 1.96 was exceeded. Nevertheless, the trial was continued. At the sixth meeting, when the O'Brien–Fleming boundary was crossed, a decision was made to terminate the trial with the final mortality curves as seen earlier in Fig. 16.5. However, it should be emphasized that crossing the boundary was not the only factor in this decision.

Flexible Group Sequential Procedures: Alpha Spending Functions

While the group sequential methods described are an important advance in data monitoring, the Beta-blocker Heart Attack Trial (BHAT) [62, 63] experience suggested two limitations. One was the need to specify the number K of planned interim analyses in advance. The second was the requirement for equal numbers of either participants or events between each analysis. This also means that the exact time of the interim analysis must be pre-specified. As indicated in the BHAT example, the numbers of deaths between analyses were not equal and exactly seven analyses of the data had been specified. If the monitoring committee had

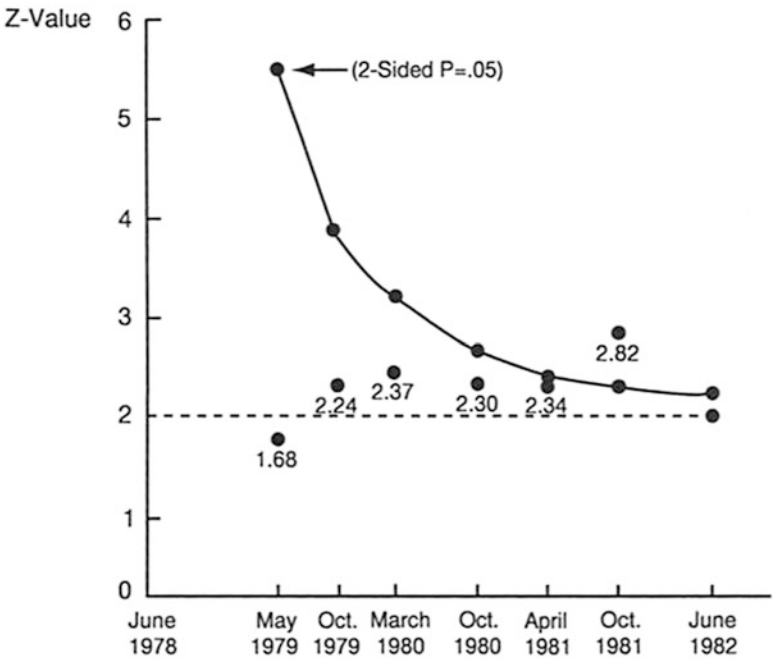

Fig. 17.2 Six interim log rank statistics plotted for the time of data monitoring committee meetings with a two-sided O'Brien-Fleming significance level boundary in the Beta-Blocker Heart Attack Trial. *Dashed line* represents $Z = 1.96$ [63]

requested an additional analysis between the fifth and sixth scheduled meetings, the O'Brien–Fleming group sequential procedure would not have directly accommodated such a modification. Yet such a request could easily have happened. In order to accommodate the unequal numbers of participants or events between analyses and the possibility of larger or fewer numbers of interim analyses than pre-specified, flexible procedures that eliminated those restrictions were developed [64–71]. The authors proposed a so-called alpha spending function which allows investigators to determine how they want to allocate or "spend" the Type I error or alpha during the course of the trial. This function guarantees that at the end of the trial, the overall Type I error will equal the prespecified value of α. As will be described, this approach is a generalization of the previous group sequential methods so that the Pocock [39] and O'Brien–Fleming [42] monitoring procedures become special cases.

We must first distinguish between calendar time and information fraction [70, 71]. The information expected from all participants at the planned end of the trial is the total information. At any particular calendar time t during the study, a certain fraction t^* of the total information is observed. That may be approximated by the fraction of participants randomized at that point, n, divided by the total number expected, N, or in survival studies, by the number of events observed

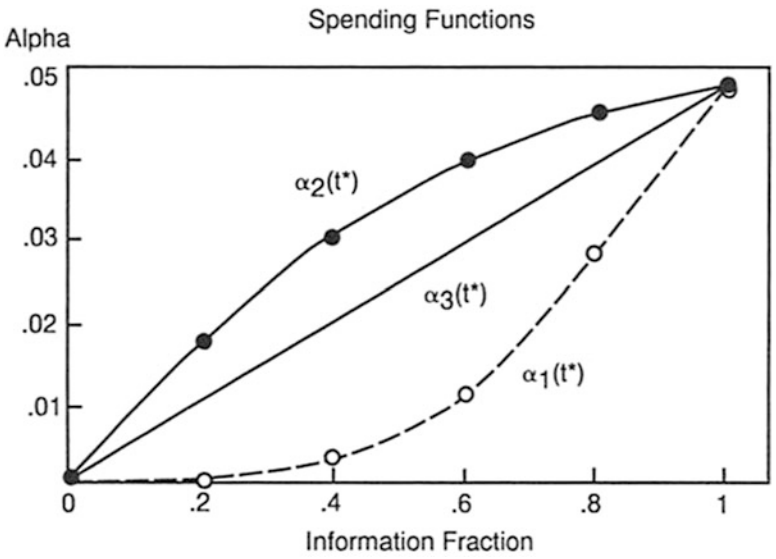

Fig. 17.3 Alpha-spending functions for K = 5, two-sided α = 0.05 at information fractions 0.2, 0.4, 0.6, 0.8, and 1.0. α1(t*) ~ O'Brien-Fleming; α2(t*) ~ Pocock; α3(t*) ~ uniform [74]

already, d, divided by the total number expected D. Thus, the value for t^* must be between 0 and 1. The information fraction is more generally defined in terms of ratio of the inverse of the variance of the test statistic at the particular interim analysis and the final analysis. The alpha spending function, $\alpha(t^*)$, determines how the prespecified α is allocated at each interim analyses as a function of the information fraction. At the beginning of a trial, $t^* = 0$ and $\alpha(t^*) = 0$, while at the end of the trial, $t^* = 1$ and $\alpha(t^*) = \alpha$. Alpha-spending functions that correspond to the Pocock and O'Brien–Fleming boundaries shown in Fig. 17.1 are indicated in Fig. 17.3 for a two-sided 0.05 α level and five interim analyses. These spending functions correspond to interim analyses at information fractions at 0.2, 0.4, 0.6, 0.8, and 1.0. However, in practice the information fractions need not be equally spaced. We chose those information fractions to indicate the connection between the earlier discussion of group sequential boundaries and the α spending function. The Pocock-type spending function allocates the alpha more rapidly than the O'Brien–Fleming type spending function. For the O'Brien–Fleming-type spending function at $t^* = 0.2$, the $\alpha(0.2)$ is less than 0.0001 which corresponds approximately to the very large critical value or boundary value of 4.56 in Fig. 17.1. At $t^* = 0.4$, the amount of α which can be spent is $\alpha(0.4) - \alpha(0.2)$ which is approximately 0.0006, corresponding to the boundary value 3.23 in Fig. 17.1. That is, the difference in $\alpha(t^*)$ at two consecutive information fractions, t^* and t^{**} where t^* is less than t^{**}, $\alpha(t^{**}) - \alpha(t^*)$, determines the boundary or critical value at t^{**}. Obtaining these critical values consecutively requires numerically integrating a distribution function similar to that for the Pocock boundary and is described elsewhere in detail [68]. Because these spending functions are only approximately equivalent to the

Pocock or O'Brien–Fleming boundaries, the actual boundary values will be similar but not exactly the same. However, the practical operational differences are important in allowing greater flexibility in the monitoring process. Programs are available for these calculations [72, 73].

Many different spending functions can be specified. The O'Brien–Fleming $\alpha_1(t^*)$ and Pocock α_2 (t^*) type spending functions are specified as follows:

$$\begin{aligned}
\alpha_1(t^*) &= 2 - 2\Phi\left(Z_{\alpha/2}/\sqrt{t^*}\right) && \sim \text{O'Brien-Fleming} \\
\alpha_2(t^*) &= \alpha \ln\left(1 + (e-1)t^*\right) && \sim \text{Pocock} \\
\alpha_3(t^*) &= \alpha\, t^{*\theta} && \text{for} \quad \theta > 0
\end{aligned}$$

The spending function $\alpha_3(t^*)$ spends alpha uniformly during the trial for $\theta = 1$, at a rate somewhat between $\alpha_1(t^*)$ and $\alpha_2(t^*)$. Other spending functions have also been defined [75, 76].

The advantage of the alpha-spending function is that neither the number nor the time of the interim analyses needs to be specified in advance. Once the particular spending function is selected, the information fractions t_1^*, t_2^*, ... determine the critical or boundary values exactly. In addition, the frequency of the interim analyses can be changed during the trial and still preserve the prespecified α level. Even if the rationale for changing the frequency is dependent on the emerging trends, the impact on the overall Type I error rate is almost negligible [77, 78]. These advantages give the spending function approach to group sequential monitoring the flexibility in analysis times that is often required in actual clinical trial settings [79]. It must be emphasized that no change of the spending function itself is permitted during the trial. Other authors have discussed additional aspects of this approach [80–82].

Applications of Group Sequential Boundaries

As indicated in the BHAT example [62, 63], the standardized logrank test can be compared to the standardized boundaries provided by the O'Brien–Fleming, Pocock, or α spending function approach. However, these group sequential methods are quite widely applicable for statistical tests. Under very general conditions, any statistic testing a single parameter from a parametric or semiparametric model has the normal or asymptotically normal distribution with independent increments of information between interim analyses which is sufficient for this approach [83, 84]. Many of the commonly used test statistics used in clinical trials have this feature. Besides logrank and other survival tests, comparisons of means, comparison of proportions [39, 85] and comparison of linear regression slopes [86–91] can be monitored using this approach. For means and proportions, the information fraction can be approximated by the ratio of the number of participants observed to the total expected. For regression slopes, the information fraction is

best determined from the ratio of the inverse of the variance of the regression slope differences computed for the current and expected final estimate [86, 90, 91]. Considerable work has extended the group sequential methodology to more general linear and nonlinear random effects models for continuous data and to repeated measure methods for categorical data [83, 84, 92]. Thus, for most of the statistical tests that would be applied to common primary outcome measures in a clinical trial setting, the flexible group sequential methods can be used directly.

If the trial continues to the scheduled termination point, a p value is often computed to indicate the extremeness of the result. If the standardized statistical test exceeds the critical value, the p value would be less than the corresponding significance level. If a trial is terminated early or continues to the end with the standardized test exceeding or crossing the boundary value, a p value can also be computed [93]. These p values cannot be the nominal p value corresponding to the standardized test statistic. They must be adjusted to account for the repeated statistical testing of the outcome measure and for the particular monitoring boundary employed. Calculation of the p value is relatively straight forward with existing software packages [72, 73].

Statistical tests of hypotheses are but one of the methods used to evaluate the results of a clinical trial. Once trials are terminated, either on schedule or earlier, confidence intervals (CIs) are often used to give some sense of the uncertainty in the estimated treatment or intervention effect. For a fixed sample study, CIs are typically constructed as

$$(\text{effect estimate}) \pm Z(\alpha)\ \text{SE}(\text{estimate})$$

where SE is the standard error of the estimate.

In the group sequential monitoring setting, this CI will be referred to as the naïve estimate since it does not take into account the sequential testing aspects. In general, construction of CIs following the termination of a clinical trial is not as straightforward [94–107], but software exists to aid in the computations [72]. The major problem with naive CIs is that they may not give proper coverage of the unknown but estimated treatment effect. That is, the CIs constructed in this way may not include the true effect with the specified frequency (e.g., 95%). For example, the width of the CI may be too narrow. Several methods have been proposed for constructing a more proper CI [94–107] by typically ordering the possible outcomes in different ways. That is, a method is needed to determine if a treatment effect at one time is either more or less extreme than a difference at another time. None of the methods proposed appear to be universally superior but the ordering originally suggested by Siegmund [104] and adopted by Tsiatis et al. [105] appears to be quite adequate in most circumstances. In this ordering, any treatment comparison statistic which exceeds the group sequential boundary at one time is considered to be more extreme than any result which exceeds the sequential boundary at a later time. While construction of CIs using this ordering of possible outcomes can break down, the cases or circumstances are almost always quite unusual and not likely to occur in practice [107]. It is also interesting that for conservative monitoring boundaries

such as the O'Brien– Fleming method, the naive CI does not perform that poorly, due primarily to the extreme early conservatism of the boundary [103]. While more exact CIs can be computed for this case, the naive estimate may still prove useful as a quick estimate to be recalculated later using the method described [105]. Pocock and Hughes [102] have suggested that the point estimate of the effect of the intervention should also be adjusted, since trials that are terminated early tend to exaggerate the size of the true treatment difference. Others have also pointed out the bias in the point estimate [96, 101]. Kim [101] suggested that an estimate of the median is less biased.

CIs can also be used in another manner in the sequential monitoring of interim data. At each interim analysis, a CI could be constructed for the parameter summarizing the intervention effect, such as differences in means, proportions, or hazard ratios. This is referred to as repeated confidence intervals (RCIs) [95, 98, 99]. If the RCI excludes a null difference, or no intervention effect, then the trial might be stopped claiming a significant effect, either beneficial or harmful. It is also possible to continue the trial unless the CI excluded not only no difference but also minimal or clinically unimportant differences. On the other hand, if all values of clinically meaningful treatment differences are ruled out or fall outside the CI, then that trial might be stopped claiming that no useful clinical effect is likely. This method is useful for non-inferiority designs as described earlier in Chap. 5. Here, as for CIs following termination, the naive CI is not appropriate. Jennison and Turnbull [98, 99] have suggested one method for RCIs that basically inverts the group sequential test. That is, the CI has the same form as the naive estimate, but the coefficient is the standardized boundary value as determined by the spending function, for example. The RCI then has the following form:

$$(\text{treatment difference}) \pm Z(k)\text{SE}(\text{difference})$$

where $Z(k)$ is the sequential boundary value at the kth interim analysis. For example, using the O'Brien–Fleming boundaries shown in Fig. 17.1, we would have a coefficient of 4.56 at $k = 1$, $t_1^* = 0.2$ and 3.23 at $k = 2$, $t_2^* = 0.4$. Used in this manner, the RCI and the sequential test of the null hypothesis will yield the same conclusions.

One particular application of the RCI is for trials whose goal is to demonstrate that two interventions or treatments are essentially equivalent, that is, have an effect that is considered to be within a specified acceptable range and might be used interchangeably. As indicated in Chap. 5, clinicians might select the cheaper, less toxic or less invasive intervention if the effects were close enough. One suggestion for "close enough" or "equivalence" would be treatments whose effects are within 20% [108, 109]. Thus, RCIs that are contained within a 20% range would suggest that the results are consistent with this working definition of equivalence. For example, if the relative risks were estimated along with a RCI, the working range of equivalence would be from 0.8 to 1.2, where large values indicate inferiority of the intervention being tested. The trial would continue as long as the upper limit of the RCI exceeded 1.2 since we would not have ruled out a treatment worsening by

20% or more. Depending on the trial and the interventions, the trial might also continue until the lower limit of the RCI was larger than 0.8, indicating no improvement by 20% or greater.

As described in Chap. 5, there is a fundamental difference between an "equivalence" design and a noninferiority design. The former is a two-sided test, with the aim of establishing a narrow range of possible differences between the new intervention and the standard, or that any difference is within a narrow range. The noninferiority design aims to establish that the new intervention is no worse than the standard by some prespecified margin. It may be that the margins in the two designs are set to the same value. From a data monitoring point of view, both of these designs are best handled by sequential CIs [99]. As data emerge, the RCI takes into consideration the event rate or variability, the repeated testing aspects, and the level of the CI. The upper and lower boundaries can address either the "equivalence" point of view or the noninferiority margin of indifference.

Asymmetric Boundaries

In most trials, the main purpose is to test whether the intervention is superior to the control. It is rarely ethical to continue a study in order to prove, at the usual levels of significance, that the intervention is harmful relative to a placebo or standard control. This point has been mentioned by authors [110, 111] who discuss methods for group sequential designs in which the hypothesis to be tested is one-sided; that is, to test whether the intervention is superior to the control. They proposed retaining the group sequential upper boundaries of methods such as Pocock, Haybittle–Peto, or O'Brien–Fleming for rejection of H_0 while suggesting various forms of a lower boundary which would imply "acceptance" of H_0. One simple approach is to set the lower boundary at an arbitrary value of Z_i, such as -1.5 or -2.0. If the test statistic goes below that value, the data may be sufficiently suggestive of a harmful effect to justify terminating the trial. This asymmetric boundary attempts to reflect the behavior or attitude of members of many monitoring committees, who recommend stopping a study once the intervention shows a strong, but non-significant, trend in an adverse direction for major events. Emerson and Fleming [112] recommend a lower boundary for acceptance of the null hypothesis which allows the upper boundary to be changed in order to preserve the Type I error exactly. Work by Gould and Pecore [113] suggests ways for early acceptance of the null hypothesis while incorporating costs as well. For new interventions, trials might well be terminated when the chances of a positive or beneficial result seem remote (discussed in the next section). However, if the intervention arm is being compared to a standard but the intervention is already in widespread use, it may be important to distinguish between lack of benefit and harm [114]. For example, if the intervention is not useful for the primary outcome, and also not harmful, it may still have benefits such as on other secondary clinical outcomes, quality of life, or fewer adverse events that would still make it a

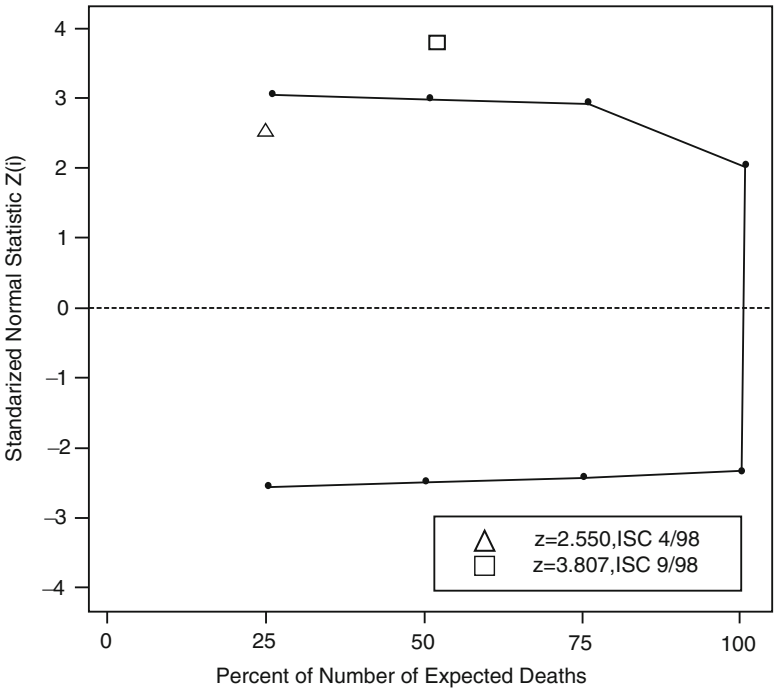

Fig. 17.4 MERIT-HF group sequential monitoring bounds for mortality [118]

therapeutic option. In such cases, a symmetric boundary for the primary outcome might be appropriate.

An example of asymmetric group sequential boundaries is provided by the Cardiac Arrhythmia Suppression Trial (CAST). Two arms of the trial (encainide and flecainide, each vs. placebo) were terminated early using a symmetric two-sided boundary, although the lower boundary for harm was described as advisory by the authors [115–117]. The third comparison (moricizine vs. placebo) continued. However, due to the experience with the encainide and flecainide arms, the lower boundary for harm was revised to be less stringent than originally, i.e., an asymmetric boundary was used [115].

MERIT-HF used a modified version of the Haybittle–Peto boundary for benefit, requiring a critical value near +3.0 and a similar but asymmetric boundary, close to a critical Z value of -2.5 for harm as shown in Fig. 17.4. In addition, at least 50% of the designed person years of exposure were to be observed before early termination could be recommended. The planned interim analyses to consider benefit were at 25, 50, and 75% of the expected target number of events. Because there was a concern that treating heart failure with a beta blocker might be harmful, the monitoring committee was required to evaluate safety on a monthly basis using the lower sequential boundary as a guide. At the 25% interim analyses, the statistic for the logrank test was +2.8, just short of the boundary for benefit. At the 50%

interim analyses, the observed logrank statistic was +3.8, clearly exceeding the sequential boundary for benefit. It also met the desired person years of exposure as plotted in Fig. 17.4. Details of this experience are described elsewhere [118]. A more detailed presentation of group sequential methods for interim analysis of clinical trials may be found in books by Jennison and Turnbull [119] and Proschan, Lan, and Wittes [120].

Curtailed Sampling and Conditional Power Procedures

During the course of monitoring accumulating data, one question often posed is whether the current trend in the data is so impressive that "acceptance" or rejection of H_0 is already determined, or at least close to being determined. If the results of the trial are such that the conclusions are known for certain, no matter what the future outcomes might be, then consideration of early termination is in order. A helpful sports analogy is a baseball team "clinching the pennant" after winning a specific game. At that time, it is known for certain who has won and who has not won the pennant or league championship, regardless of the outcome of the remaining games. Playing the remaining games is done for reasons (e.g., fiscal) other than deciding the winner. This idea has been developed for clinical trials and is often referred to as deterministic curtailed sampling. It should be noted that group sequential methods focus on existing data while curtailed sampling in addition considers the data which have not yet been observed.

Alling [121, 122] developed a closely related approach when he considered the early stopping question and compared the survival experience in two groups. He used the Wilcoxon test for two samples, a frequently used non-parametric test which ranks survival times and which is the basis for one of the primary survival analysis techniques. Alling's method allows stopping decisions to be based on data available during the trial. The trial would be terminated if future data could not change the final conclusion about the null hypothesis. The method is applicable whether all participants are entered at the same time or recruitment occurs over a longer period of time. However, when the average time to the event is short relative to the time needed to enroll participants, the method is of limited value. The repeated testing problem is irrelevant, because any decision to reject the null hypothesis is based on what the significance test will be at the end of the study. Therefore, frequent use of this procedure during the trial causes no problem with regard to significance level and power.

Many clinical trials with survival time as a response variable have observations that are censored; that is, participants are followed for some length of time and then at some point, no further information about the participant is known or collected. Halperin and Ware [123] extended the method of Alling to the case of censored data, using the Wilcoxon rank statistic. With this method, early termination is particularly likely when the null hypothesis is true or when the expected difference between groups is large. The method is shown to be more effective for small sample

sizes than for large studies. The Alling approach to early stopping has also been applied to another commonly used test, the Mantel–Haenszel statistic. However, the Wilcoxon statistic appears to have better early stopping properties than the Mantel–Haenszel statistic.

A deterministic curtailed procedure has been developed [124] for comparing the means of two bounded random variables using the two sample t-test. It assumes that the response must be between two values, A and B $(A < B)$. An approximate solution is an extreme case approach. First, all the estimated remaining responses in one group are given the maximum favorable outcome and all the remaining responses in the other take on the worst response. The statistic is then computed. Next, the responses are assigned in the opposite way and a second statistic is computed. If neither of these two extreme results alters the conclusion, no additional data are necessary for testing the hypothesis. While this deterministic curtailed approach provides an answer to an interesting question, the requirement for absolute certainty results in a very conservative test and allows little opportunity for early termination.

In some clinical trials, the final outcome may not be absolutely certain, but almost so. To use the baseball analogy again, a first place team may not have clinched the pennant but is so many games in front of the second place team that it is highly unlikely that it will not, in fact, end up the winner. Another team may be so far behind that it cannot "realistically" catch up. In clinical trials, this idea is often referred to as stochastic curtailed sampling or conditional power. It is identical to the concept of conditional power discussed in the section on extending a trial.

One of the earliest applications of the concept of conditional power was in the CDP [1, 125]. In this trial, several treatment arms for evaluating cholesterol lowering drugs produced negative trends in the interim results. Through simulation, the probability of achieving a positive or beneficial result was calculated given the observed data at the time of the interim analysis. Unconditional power is the probability at the beginning of the trial of achieving a statistically significant result at a prespecified alpha level and with a prespecified alternative treatment effect. Ideally, trials should be designed with a power of 0.80–0.90 or higher. However, once data begin to accumulate, the probability of attaining a significant result increases or decreases with emerging positive or negative trends. Calculating the probability of rejecting the null hypothesis of no effect once some data are available is conditional power.

Lan et al. [126] considered the effect of stochastic curtailed or conditional power procedures on Type I and Type II error rates. If the null hypothesis, H_0, is tested at time t using a statistic, $S(t)$, then at the scheduled end of a trial at time T, the statistic would be $S(T)$. Two cases are considered. First, suppose a trend in favor of rejecting H_0 is observed at time $t < T$, with intervention doing better than control. One then computes the conditional probability, γ_0 of rejecting H_0 at time T; that is, $S(T) > Z_\alpha$, assuming H_0 to be true and given the current data, $S(t)$. If this probability is sufficiently large, one might argue that the favorable trend is not going to disappear. Second, suppose a negative trend or data consistent with the null hypothesis of no difference, at some point t. Then, one computes the conditional probability, γ_1, of

rejecting H_0 at the end of the trial, time T, given that some alternative H_1 is true, for a sample of reasonable alternatives. This essentially asks how large the true effect must be before the current "negative" trend is likely to be reversed. If the probability of a trend reversal is highly unlikely for a realistic range of alternative hypotheses, trial termination might be considered.

Because there is a small probability that the results will change, a slightly greater risk of a Type I or Type II error rate will exist than would be if the trial continued to the scheduled end [127]. However, it has been shown that the Type I error is bounded very conservatively by α/γ_0 and the Type II error by β/γ_1. For example, if the probability of rejecting the null hypothesis, given the existing data were 0.85, then the actual Type I error would be no more than 0.05/0.85 or 0.059, instead of 0.05. The actual upper limit is considerably closer to 0.05, but that calculation requires computer simulation. Calculation of these probabilities is relatively straightforward and the details have been described by Lan and Wittes [128]. A summary of these methods, using the approach of DeMets [74], follows.

Let $Z(t)$ represent the standardized statistic at information fraction t. The information fraction may be defined, for example, as the proportion of expected participants or events observed so far. The conditional power, CP, for some alternative intervention effect θ, using a critical value of Z_α for a Type I error of alpha, can be calculated as

$$P\left[Z(1) \geq Z_\alpha \middle| Z(t), \theta\right] = 1 - \Phi\left\{\left|Z_\alpha - Z(t)\sqrt{t} - \theta(1-t)\right|/\sqrt{1-t}\right\}$$

where $\theta = E(Z(t=1))$, the expected value of the test statistic at the full completion of the trial.

The alternative θ is defined for various outcomes as follows for:

1. Survival outcome (D = total events)

$$\theta = \sqrt{D/4}\,\mathrm{Log}(\lambda_C/\lambda_T)$$

λ_C and λ_T are the hazard rates in the control and intervention arms, respectively.

2. Binomial outcome ($2n$ = N, n/arm or N = total sample size)

$$\theta = \frac{P_C - P_T}{\sqrt{2\overline{p}(1-\overline{p})/(n/2)}} = \frac{(P_C - P_T)\sqrt{N/4}}{\sqrt{\overline{p}(1-\overline{p})}}$$
$$= 1/2\,\frac{(P_C - P_T)\sqrt{N}}{\sqrt{\overline{pq}}}$$

where P_C and P_T are the event rates in the control arm and intervention arm respectively and $\overline{p}$ is the common event rate.

3. Continuous outcome (means) (N = total sample size)

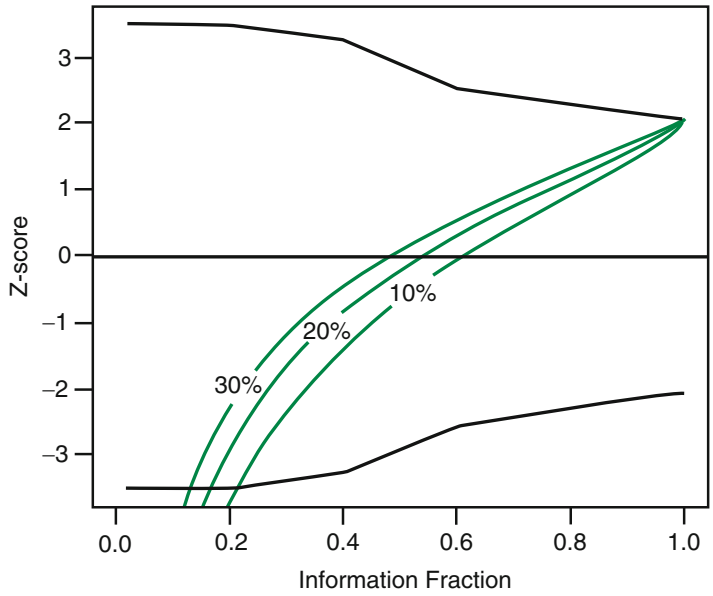

Fig. 17.5 Conditional power boundaries: outer boundaries represent symmetric O'Brien-Fleming type sequential boundaries ($\alpha = 0.05$). Three lower boundaries represent boundaries for 10%, 20% and 30% conditional power to achieve a significant ($P < 0.05$) result of the trial conclusion [74]

$$\theta = \left(\frac{\mu_C - \mu_T}{\sigma}\right)\sqrt{N/4}$$
$$= 1/2\left(\frac{\mu_C - \mu_T}{\sigma}\right)\sqrt{N}$$

where μ_C and μ_T are the mean response levels for the control and the intervention arms, respectively, and σ is the common standard deviation.

If we specify a particular value of the conditional power as γ, then a boundary can also be produced which would indicate that if the test statistic fell below that, the chance of finding a significant result at the end of the trial is less than γ [127]. For example, in Fig. 17.5 the lower futility boundary is based on a specified conditional power γ, ranging from 10 to 30% that might be used to claim futility of finding a positive beneficial claim at the end of the trial. For example, if the standardized statistic crosses that 20% lower boundary, the conditional power for a beneficial result at the end of the trial is less than 0.20 for the specified alternative.

Conditional power calculations are done for a specific alternative but in practice, a monitoring committee would likely consider a range of possibilities. These specified alternatives may range between the null hypothesis of no effect and the prespecified design based alternative treatment effect. In some cases, a monitoring committee may consider even more extreme beneficial effects to determine just how much more effective the treatment would have to be to raise the conditional

power to desired levels. These conditional power results can be summarized in a table or a graph, and then monitoring committee members can assess whether they believe recovery from a substantial negative trend is likely.

Conditional power calculations were utilized in the Vesnarinone in Heart Failure Trial (VEST) [129]. In Table 17.1, the test statistics for the logrank test are provided for the information fractions at a series of monitoring committee meetings. Table 17.2 provides conditional power for VEST at three of the interim analyses. A range of intervention effects was used including the beneficial effect (hazard rate less than 1) seen in a previous vesnarinone trial to the observed negative trend (hazard rates of 1.3 and 1.5). It is clear that the conditional power for a beneficial effect was very low by the midpoint of this trial for a null effect or worse. In fact, the conditional power was not encouraging even for the original assumed effect. As described by DeMets et al. [114] the trial continued beyond this point due to the existence of a previous trial that indicated a large reduction in mortality, rather than the harmful effect observed in VEST.

The Beta-Blocker Heart Attack Trial [62, 63] made considerable use of this approach. As discussed, the interim results were impressive with 1 year of follow-up still remaining. One question posed was whether the strong favorable trend ($Z = 2.82$) could be lost during that year. The probability of rejecting H_0 at the scheduled end of the trial, given the existing trend (γ_0), was approximately 0.90. This meant that the false positive or Type I error was no more than $\alpha/\gamma_0 = 0.05/0.90$ or 0.056.

Table 17.1 Accumulating results for the Vesnarinone in Heart Failure Trial (VEST) [129]

Information fraction	Log-rank Z-value (high dose)
0.43	+0.99
0.19	−0.25
0.34	−0.23
0.50	−2.04
0.60	−2.32
0.67	−2.50
0.84	−2.22
0.20	−2.43
0.95	−2.71
1.0	−2.41

Table 17.2 Conditional power for the Vesnarinone in Heart Failure Trial (VEST) [129]

	Information fraction		
RR	0.50	0.67	0.84
0.50	0.46	<0.01	<0.01
0.70	0.03	<0.01	<0.01
1.0	<0.01	<0.01	<0.01
1.3	<0.01	<0.01	<0.01
1.5	<0.01	<0.01	<0.01

RR = relative risk

Other Approaches

Other techniques for interim analysis of accumulating data have also received attention. These include binomial sampling strategies [15], decision theoretic models [130], and likelihood or Bayesian methods [131–140]. Bayesian methods require specifying a prior probability on the possible values of the unknown parameter. The experiment is performed and based on the data obtained, the prior probability is adjusted. If the adjustment is large enough, the investigator may change his opinion (i.e., his prior belief). Spiegelhalter et al. [139] and Freedman et al. [135] have implemented Bayesian methods that have frequentist properties very similar to boundaries of either the Pocock or O'Brien–Fleming type. It is somewhat reassuring that two methodologies, even from a different theoretical framework, can provide similar monitoring procedures. While the Bayesian view is critical of the hypothesis testing methods because of the arbitrariness involved, the Bayesian approach is perhaps hampered mostly by the requirement that the investigator formally specify a prior probability. However, if a person during the decision-making process uses all of the factors and methods discussed in this chapter, a Bayesian approach is involved, although in a very informal way.

One Bayesian method to assess futility that has been used extensively is referred to as predictive power and is related to the concept of conditional power. In this case, the series of possible alternative intervention effects, θ, are represented by a prior distribution for θ, distributing the probability across the alternatives. The prior probability distribution can be modified by the current trend to give an updated posterior for θ. The conditional power is calculated as before for a specific value of θ. Then a predictive or "average" power is calculated by integrating the conditional power over the posterior distribution for θ:

$$p\left(X_f \in R \middle| x_0\right) = \int p\left(X_f \in R \middle| \theta\right) p\left(\theta \middle| x_0\right) d\theta$$

This can then be utilized by the monitoring committee to assess whether the trial is still viable, as was computed for the interim analyses conducted in VEST [129] as shown in Table 17.3. In this case, the prior was taken from an earlier trial of vesnarinone where the observed reduction in mortality was over 60% (relative risk = 0.40). For these calculations, the prior was first set at the point estimate of the hazard ratio equal to 0.40. Using this approach, it is clear that VEST would not likely have shown a benefit at the end of the trial.

We have stated that the monitoring committee should be aware of all the relevant information in the use of the intervention which existed before the trial started and which emerges during the course of a trial. Some have argued that all of this information should be pooled or incorporated and updated sequentially in a formal statistical manner [141]. This is referred to as cumulative meta-analysis (see Chap. 18). We do not generally support cumulative or sequential meta-analysis as a primary approach for monitoring a trial. We believe that the results of the ongoing

Table 17.3 Predictive probability for the Vesnarinone in Heart Failure Trial (VEST) [129]

		Probability
Date	T[a]	Hazard rate = 0.40
2/7/96	0.50	0.28
3/7/96	0.60	0.18
4/10/96	0.67	<0.0001
5/19/96	0.84	<0.0001
6/26/96	0.90	<0.0001

[a]T = information fraction

trial should be first presented alone, in detail, including baseline comparisons, primary and secondary outcomes, adverse events and relevant laboratory data (see Chap. 16). As supportive evidence for continuation or termination, results or other analysis from external completed trials may be used, including a pooled analysis of all available external data.

Trend Adaptive Designs and Sample Size Adjustments

Sample size adjustments based on overall event rates or outcome variability, without knowledge of interim trends, have long been performed to regain trial power with no issues regarding impact on Type I error or other design concerns. However, while sample size adjustments based on comparing emerging trends in the intervention and control groups were initially discouraged, statistical methodology now allows trialists to adjust the sample size and maintain the Type I error while regaining power [142–164]. It is possible to have a statistically efficient or nearly efficient design if the adaptations are prespecified [154]. While multiple adjustments over the course of follow-up are possible, the biggest gain comes from a single adaptive adjustment.

These methods must be implemented by some individual or party that is aware of the emerging trend. In general, we do not recommend that the monitoring committee perform this function because it may be aware of other factors that would mitigate any sample size increase but cannot share those issues with the trial investigators or sponsors. This can present an awkward if not an ethical dilemma for the monitoring committee. Rather, someone who only knows the emerging trend should make the sample size adjustment recommendation to the investigators. Whatever trend adaptive method is used must also take into account the final analyses as discussed briefly in Chap. 18, because it can affect the final critical value. We will briefly describe a few of these methods [145, 147, 159].

As proposed by Cui et al. [146, 147] for adaptive adjustments in a group sequential setting, suppose we measure an outcome variable denoted as X where X has a $N(0,1)$ distribution and n is current sample size, N_0 is initial total sample size, N is new target sample size, θ is hypothesized intervention effect, and t is n/N_0. In this case, we can have an estimate of the intervention effect and a test statistic

based on n observations.

$$\hat{\theta} = \sum_{i}^{n} x_i / n$$
$$z^{(n)} = \sum_{i}^{n} x_i / \sqrt{n}$$

We then compute a revised sample size N based on the current trend, assuming the same initial Type I error and desired power. A new test statistic is defined that combines the already observed data and the yet to be obtained data.

$$Z_W^{(N)} = \sqrt{t} Z^{(n)} + \sqrt{1-t}(N-n)^{-\frac{1}{2}} \sum_{n+1}^{N} x_{\mathrm{i}}$$

In this setting, we would reject the null hypothesis H_0 of no treatment effect if $Z_W^{(N)} > Z_\alpha$. This revised test statistic controls the Type I error at the desired level α. However, less weight is assigned to the new or additional observations, yielding what is referred to as a weighted Z statistic. That is, the weight given to each trial participant prior to any sample size adjustment is greater than weight given to participants after the adjustment, violating a "one participant—one vote" principle. This discounting may not be acceptable for scientific and ethical reasons [144, 164].

Other approaches have also been proposed. A modification proposed by Chen et al. [145] requires that both the weighted and un-weighted test statistics exceed the standard critical value.

$$Z^{(N)} \quad \text{and} \quad Z_W^{(N)} > Z_\alpha$$

In this case, the Type I error $< \alpha$ and there is no loss of power. Another approach, an adjusted p value method, proposed by Proschan and colleagues [159, 160] requires a "promising" p value before allowing an increase in sample size. However, this approach requires stopping if the first stage p value is not promising. It also requires a larger critical value at the second stage to control the Type I error. As an example, consider a one-sided significance level $a = 0.05$, which would ordinarily have a critical value of 1.645 for the final test statistic. In this case the promising p value, p', and the final critical values, Z', are as follows, regardless of the sample size in the second stage:

p':	0.10	0.15	0.20	0.25	0.50
Z':	1.77	1.82	1.85	1.875	1.95

This simple method will control the Type I error but in fact may make Type I error substantially less than 0.05. A method can be developed to obtain an exact

Type I error as a function of $Z(t)$ and the adjusted sample size N, using a conditional power type calculation [127] as described below.

Conditional power, CP, is a useful calculation to assess the likelihood of exceeding a critical value at the scheduled end of a trial, given the current data or value of the interim test statistic and making assumptions about the future intervention effect as described earlier in this chapter [67, 126, 128]. The computation of conditional power in this case is relatively simple. Let θ be a function of the intervention affect, as described earlier, and then

$$\begin{aligned} CP(Z(t),\theta) &= P\left[Z(T) \geq Z_\alpha \middle| Z(t),\theta\right) \\ &= 1 - \Phi\left\{\left|Z_\alpha - Z(t)\sqrt{t} - \theta(1-t)\right| \Big/ \sqrt{(1-t)}\right\} \end{aligned}$$

Applying the idea of conditional power to the trend adaptive design, we can define an algorithm to adjust the sample size and still control the Type I error [146]. For example,

Let $\Delta =$ observed effect and $\delta =$ assumed effect. If we observe that for $\theta(\Delta)$ as a function of the observed effect Δ, and θ (δ) as a function of the assumed δ, then if

$$\begin{aligned} \text{CP}(Z(t),\ \theta(\Delta)) &> 1.2CP\left(Z(t),\theta(\delta), \quad \text{decrease N}\right. \\ \text{CP}(Z(t),\ \theta(\Delta)) &< 0.8CP\left(Z(t),\theta(\delta), \quad \text{increase N}\right. \end{aligned}$$

where N is the final targeted sample size. The properties of this procedure have not been well investigated but the idea is related to other conditional power approaches [153]. These conditional power procedures adjust the sample size if the computed conditional power for the current trend is marginal, with only a trivial impact on Type I error. For example, define a lower limit (c_ℓ) and an upper limit (c_u) such that for the current trend $\theta(\Delta)$:

if $\text{CP}(Z(t),\theta(\Delta)) < c_\ell$, then terminate for futility and accept the null (required),
if $\text{CP}(Z(t),\theta(\Delta)) > c_\text{u}$, then continue with no change in sample size, or
if $c_\ell < \text{CP}(Z(t),\theta(\Delta)) < c_\text{u}$, then increase sample size from N_0, to N to get conditional power to the desired level.

Chen et al. [145] suggested a modest alternative. If the conditional power is 50% or larger, then increase the sample size to get the desired power. An upper cap is typically placed on the size of the increase in sample size. Increase N_0 if the interim result is "promising," defined as conditional power >50% for the current trend but the increase in N_0 cannot be greater than 1.75-fold. Under these conditions, Type I error is not increased and there is no practical loss in power. This approach is one that we favor since it is simple to implement, easy to understand and preserves the design characteristics.

Adaptive designs have appeal because the assumptions made during protocol development often fail to hold precisely for the implemented trial, making adjustments useful or even necessary for the study to succeed. However, adaptive designs also rely on assumptions which prove to be unmet in practice, so that theoretical

gains are not necessarily realized. For example, it is often found that the observed event rate is less than expected, or the intervention effect not as great as had been assumed. Tsiatis and Mehta [163] have provided conditions under which a properly designed group sequential trial is more efficient than these adaptive designs, though Mehta has also argued that factors such as allocation of financial and participant resources may be as important as statistical efficiency [157]. In any case, a clear need exists for adaptive designs, including trend adaptive designs. We are fortunate that technical advances have been made through several new methods. Research continues on finding methods which can applied to different trial settings [143, 150–152, 154–158, 161, 164].

Perhaps the largest challenge is how to implement the trend adaptive design without introducing bias or leaving the door open for bias. If one utilizes one of the described trend adaptive designs, anyone who knows the details of the method can "reverse engineer" the implementation and obtain a reasonable estimate of what the current trend ($Z(t)$) must have been to generate the adjusted sample size (N). Given that these trend adaptive designs have as yet not been widely used, there is not enough experience to recommend what can be done to best minimize bias. However, as suggested earlier, a third party who knows only the emerging trend and none of the other secondary or safety data is probably best suited to make these calculations and provide them to the investigators.

References

1. Canner PL. Practical Aspects of Decision-Making In Clinical Trials—The Coronary Drug Project as a Case-Study. *Control Clin Trials* 1981;1:363–376.
2. DeMets DL. Data monitoring and sequential analysis—An academic perspective. *J Acquir Immune Defic Syndr* 1990;3:S124–S133.
3. DeMets DL, Furberg C, Friedman LM. Data monitoring in clinical trials: a case studies approach. New York, NY, Springer, 2006.
4. Ellenberg SS, Fleming TR, DeMets DL. Data Monitoring Committees in Clinical Trials: A Practical Perspective. Wiley, 2003.
5. Fisher MR, Roecker EB, DeMets DL. The role of an independent statistical analysis center in the industry-modified National Institutes of Health model. *Drug Inf J* 2001;35:115–129.
6. Fleming TR, DeMets DL. Monitoring of clinical trials: issues and recommendations. *Control Clin Trials* 1993;14:183–197.
7. Anscombe FJ. Sequential medical trials. *J Am Stat Assoc* 1963;58:365–383.
8. Armitage P. Restricted sequential procedures. *Biometrika* 1957;9–26.
9. Armitage P, McPherson CK, Rowe BC. Repeated Significance Tests on Accumulating Data. *J R Stat Soc Ser A* 1969;132:235–244.
10. Bross I Sequential medical plans. *Biometrics* 1952;8:188–205.
11. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *Am Stat* 1966;20:18–23.
12. DeMets DL, Lan KKG. An Overview of Sequential-Methods and Their Application in Clinical-Trials. *Commun Stat Theory Methods* 1984;13:2315–2338.
13. Robbins H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 1952;58:527–535.

14. Robbins H. Statistical methods related to the law of the iterated logarithm. *Ann Math Stat* 1970;1397–1409.
15. Simon R, Weiss GH, Hoel DG. Sequential Analysis of Binomial Clinical Trials. *Biometrika* 1975;62:195–200.
16. Wald A. Sequential Analysis. Dover Publications, 2013.
17. Whitehead J, Stratton I. Group Sequential Clinical Trials with Triangular Continuation Regions. *Biometrics* 1983;39:227–236.
18. Whitehead J. The Design and Analysis of Sequential Clinical Trials. Wiley, 1997.
19. Whitehead J, Jones D. The analysis of sequential clinical trials. *Biometrika* 1979;66:443–452.
20. Armitage P. Sequential medical trials, ed 2. New York, Wiley, 1975.
21. Silverman WA, Agate FJ, Fertig JW. A sequential trial of the nonthermal effect of atmospheric humidity on survival of newborn infants of low birth weight. *Pediatrics* 1963;31:719–724.
22. Truelove SC, Watkinson G, Draper G. Comparison of corticosteroid and sulphasalazine therapy in ulcerative colitis. *Br Med J* 1962;2:1708.
23. Freireich EJ, Gehan E, Frei E, et al. The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood* 1963;21:699–716.
24. McPherson CK, Armitage P. Repeated significance tests on accumulating data when the null hypothesis is not true. *J R Stat Soc Ser A* 1971;15–25.
25. Chatterjee SK, Sen PK. Nonparametric testing under progressive censoring. *Calcutta Statist Assoc Bull* 1973;22:13–50.
26. Dambrosia JM, Greenhouse SW. Early stopping for sequential restricted tests of binomial distributions. *Biometrics* 1983;695–710.
27. Whitehead J, Jones DR, Ellis SH. The analysis of a sequential clinical trial for the comparison of two lung cancer treatments. *Statist Med* 1983;2:183–190.
28. Breslow NE, Haug C. Sequential comparison of exponential survival curves. *J Am Stat Assoc* 1972;67:691–697.
29. Canner PL. Monitoring treatment differences in long-term clinical trials. *Biometrics* 1977;603–615.
30. Davis CE. A two sample Wilcoxon test for progressively censored data. *Commun Stat Theory Methods* 1978;7:389–398.
31. Joe H, Koziol JA, Petkau AJ. Comparison of procedures for testing the equality of survival distributions. *Biometrics* 1981;327–340.
32. Jones D, Whitehead J. Sequential forms of the log rank and modified Wilcoxon tests for censored data. *Biometrika* 1979;66:105–113.
33. Koziol JA, Petkau AJ. Sequential testing of the equality of two survival distributions using the modified Savage statistic. *Biometrika* 1978;65:615–623.
34. Muenz LR, Green SB, Byar DP. Applications of the Mantel-Haenszel statistic to the comparison of survival distributions. *Biometrics* 1977;617–626.
35. Nagelkerke NJD, Hart AAM. The sequential comparison of survival curves. *Biometrika* 1980;67:247–249.
36. Sellke T, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika* 1983;70:315–326.
37. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol* 1971;44:793–797.
38. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976;34:585.
39. Pocock SJ. Group Sequential Methods in Design and Analysis of Clinical-Trials. *Biometrika* 1977;64:191–200.
40. Pocock SJ. Size of cancer clinical trials and stopping rules. *Br J Cancer* 1978;38:757.

41. Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics* 1982;38:153–162.
42. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549–556.
43. DeMets DL. Practical aspects in data monitoring: a brief review. *Stat Med* 1987;6:753-760.
44. Emerson SS, Fleming TR. Interim analyses in clinical trials. *Oncology (Williston Park, NY)* 1990;4:126.
45. Fleming TR, Watelet LF. Approaches to monitoring clinical trials. *J Natl Cancer Inst* 1989;81:188–193.
46. Freedman LS, Lowe D, Macaskill P. Stopping rules for clinical trials. *Statist Med* 1983;2:167–174.
47. Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: a review and commentary. *Stat Sci* 1990;299–317.
48. Gail MH, DeMets DL, Slud EV. Simulation studies on increments of the two-sample logrank score test for survival time data, with application to group sequential boundaries. Lecture Notes-Monograph Series 1982;2:287–301.
49. Harrington DP, Fleming TR, Green SJ. Procedures for serial testing in censored survival data; in Crowley J, Johnson RA, Gupta SS (eds): Survival Analysis. Hayward, CA, Institute of Mathematical Statistics, 1982, pp 269–286.
50. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer chemotherapy reports Part 1 1966;50:163–170.
51. Tsiatis AA. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika* 1981;68:311–315.
52. Tsiatis AA. Group sequential methods for survival analysis with staggered entry. Lecture Notes-Monograph Series 1982;2:257–268.
53. Tsiatis AA. Repeated Significance Testing for A General-Class of Statistics Used in Censored Survival Analysis. *J Am Stat Assoc*1982;77:855–861.
54. Gehan EA. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika* 6-1-1965;52:203–223.
55. Slud E, Wei LJ. Two-Sample Repeated Significance Tests Based on the Modified Wilcoxon Statistic. *J Am Stat Assoc* 1982;77:862–868.
56. Peto R, Peto J: Asymptotically Efficient Rank Invariant Test Procedures. *J R Stat Soc Ser A* 1972;135:185–207.
57. DeMets DL, Gail MH. Use of logrank tests and group sequential methods at fixed calendar times. *Biometrics* 1985;41:1039–1044.
58. George SL. Sequential Methods Based on the Boundaries Approach for the Clinical Comparison Of Survival Times-Discussion. Statistics in Medicine 13[13-14], 1369–1370. 1994, John Wiley & Sons Ltd.
59. Kim K, Tsiatis AA. Study Duration for Clinical-Trials with Survival Response and Early Stopping Rule. *Biometrics* 1990;46:81–92.
60. Kim K. Study duration for group sequential clinical trials with censored survival data adjusting for stratification. *Statist Med* 1992;11:1477–1488.
61. Whitehead J: Sequential methods based on the boundaries approach for the clinical comparison of survival times. *Statist Med* 1994;13:1357–1368.
62. Beta-Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction: I. mortality results. *JAMA* 1982;247:1707–1714.
63. DeMets DL, Hardy R, Friedman LM, Gordon Lan KK. Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Control Clin Trials* 1984;5:362–372.
64. DeMets DL, Lan KK. Interim analysis: the alpha spending function approach. *Stat Med* 1994;13:1341–1352.
65. Kim K, DeMets DL: Design and Analysis of Group Sequential Tests Based on the Type I Error Spending Rate Function. *Biometrika* 1987;74:149–154.

66. Lan KKG, Rosenberger WF, Lachin JM. Use of spending functions for occasional or continuous monitoring of data in clinical trials. *Stat Med* 1993;12:2219–2231.
67. Lan KKG, Zucker DM. Sequential monitoring of clinical trials: the role of information and Brownian motion. *Stat Med* 1993;12:753–765.
68. Lan KKG, DeMets DL. Discrete Sequential Boundaries for Clinical-Trials. *Biometrika* 1983;70:659–663.
69. Lan KKG, DeMets DL, Halperin M. More Flexible Sequential and Non-Sequential Designs in Long-Term Clinical-Trials. *Commun Stat Theory Methods* 1984;13:2339–2353.
70. Lan KKG, Reboussin DM, DeMets DL. Information and Information Fractions for Design and Sequential Monitoring of Clinical-Trials. *Commun Stat Theory Methods* 1994;23:403–420.
71. Lan KKG, DeMets D. Group sequential procedures: calendar versus information time. *Stat Med* 1989;8:1191–1198.
72. Reboussin DM, DeMets DL, Kim K, Lan KKG. Lan-DeMets Method—Statistical Programs for Clinical Trials. [2.1]. 11–17–2003.
73. Reboussin DM, DeMets DL, Kim K, Lan KKG. Computations for group sequential boundaries using the Lan-DeMets spending function method. *Control Clin Trials* 2000;21:190–207.
74. DeMets DL. Futility approaches to interim monitoring by data monitoring committees. *Clin Trials* 2006;3:522–529.
75. Hwang IK, Shih WJ, De Cani JS. Group sequential designs using a family of type I error probability spending functions. *Statist Med* 1990;9:1439–1445.
76. Wang SK, Tsiatis AA: Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987;193–199.
77. Lan KKG, DeMets DL. Changing frequency of interim analysis in sequential monitoring. *Biometrics* 1989;45:1017–1020.
78. Proschan MA, Follmann DA, Waclawiw MA. Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics* 1992;1131–1143.
79. Geller NL. Discussion of "Interim analysis: the alpha spending approach". *Statist Med* 1994;13:1353–1356.
80. Falissard B, Lellouch J. A new procedure for group sequential analysis in clinical trials. *Biometrics* 1992;373–388.
81. Lan KKG, Lachin JM. Implementation of group sequential logrank tests in a maximum duration trial. *Biometrics* 1990;46:759–770.
82. Li ZQ, Geller NL. On the Choice of Times for Data Analysis in Group Sequential Clinical Trials. *Biometrics* 1991;47:745–750.
83. Jennison C, Turnbull BW. Group-sequential analysis incorporating covariate information. *J Am Stat Assoc* 1997;92:1330–1341.
84. Scharfstein DO, Tsiatis AA, Robins JM. Semiparametric Efficiency and Its Implication on the Design and Analysis of Group-Sequential Studies. *J Am Stat Assoc* 1997;92:1342–1350.
85. Kim K, DeMets DL. Sample size determination for group sequential clinical trials with immediate response. *Stat Med* 1992;11:1391–1399.
86. Lee JW, DeMets DL. Sequential Comparison of Changes with Repeated Measurements Data. *J Am Stat Assoc* 1991;86:757–762.
87. Lee JW, DeMets DL. Sequential Rank-Tests with Repeated Measurements in Clinical-Trials. *J Am Stat Assoc* 1992;87:136–142.
88. Lee JW. Group sequential testing in clinical trials with multivariate observations: a review. *Statist Med* 1994;13:101–111.
89. Su JQ, Lachin JM. Group Sequential Distribution-Free Methods for the Analysis of Multivariate Observations. *Biometrics* 1992;48:1033–1042.
90. Wei LJ, Su JQ, Lachin JM. Interim Analyses with Repeated Measurements in A Sequential Clinical-Trial. *Biometrika* 1990;77:359–364.
91. Wu MC, Lan G KK. Sequential Monitoring for Comparison of Changes in a Response Variable in Clinical Studies. *Biometrics* 1992;48:765–779.

92. Gange SJ, DeMets DL. Sequential monitoring of clinical trials with correlated responses. *Biometrika* 1996;83:157–167.
93. Fairbanks K, Madsen R. P values for tests using a repeated significance test design. *Biometrika* 1982;69:69–74.
94. Chang MN, O'Brien PC. Confidence intervals following group sequential tests. *Control Clin Trials* 1986;7:18–26.
95. DeMets DL, Lan KKG. Discussion of: Interim analyses: The repeated confidence interval approach by C. Jennison and BW Turnbull. *J R Stat Soc Series B Stat Methodol* 1989;51:344.
96. Emerson SS, Fleming TR: Parameter Estimation Following Group Sequential Hypothesis Testing. *Biometrika* 1990;77:875–892.
97. Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Stat Med* 1988;7:1231–1242.
98. Jennison C, Turnbull BW. Repeated Confidence-Intervals for Group Sequential Clinical-Trials. *Control Clin Trials* 1984;5:33–45.
99. Jennison C, Turnbull BW. Interim Analyses: The Repeated Confidence Interval Approach. *J R Stat Soc Series B Stat Methodol* 1989;51:305–361.
100. Kim K, DeMets DL. Confidence Intervals Following Group Sequential Tests in Clinical Trials. *Biometrics* 12-1-1987;43:857–864.
101. Kim K. Point Estimation Following Group Sequential Tests. *Biometrics* 1989;45:613–617.
102. Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. *Control Clin Trials* 1989;10:209S–221S.
103. Rosner GL, Tsiatis AA. Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* 1988;75:723–729.
104. Siegmund D. Estimation following sequential tests. *Biometrika* 1978;65:341–349.
105. Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics* 1984;797–803.
106. Whitehead J: On the bias of maximum likelihood estimation following a sequential test. *Biometrika* 1986;73:573–581.
107. Whitehead J, Facey KM. Analysis after a sequential trial: A comparison of orderings of the sample space. Joint Society for Clinical Trials/International Society for Clinical Biostatistics, Brussels 1991.
108. Fleming TR. Treatment evaluation in active control studies. *Cancer Treat Rep* 1987;71:1061–1065.
109. Fleming TR. Evaluation of active control trials in AIDS. *J Acquir Immune Defic Syndr* 1990;3:S82–S87.
110. DeMets DL, Ware JH. Group Sequential Methods for Clinical Trials with A One-Sided Hypothesis. *Biometrika* 1980;67:651–660.
111. DeMets DL, Ware JH. Asymmetric Group Sequential Boundaries for Monitoring Clinical-Trials. *Biometrika* 1982;69:661–663.
112. Emerson SS, Fleming TR. Symmetric Group Sequential Test Designs. *Biometrics* 1989;45:905–923.
113. Gould AL, Pecore VJ. Group sequential methods for clinical trials allowing early acceptance of Ho and incorporating costs. *Biometrika* 1982;69:75–80.
114. DeMets DL, Pocock SJ, Julian DG. The agonising negative trend in monitoring of clinical trials. *Lancet* 1999;354:1983–1988.
115. Cardiac Arrhythmia Suppression Trial Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med* 1992;327:227–233.
116. Friedman LM, Bristow JD, Hallstrom A, et al. Data monitoring in the cardiac arrhythmia suppression trial. Online Journal of Current Clinical Trials 7–31–1993;79.
117. Pawitan Y, Hallstrom A. Statistical interim monitoring of the cardiac arrhythmia suppression trial. *Statist Med* 1990;9:1081–1090.

118. Feyzi J, Julian DG, Wikstrand J, Wedel H. Data monitoring experience in the Metoprolol CR/XL randomized intervention trial in chronic heart failure: Potentially high-risk treatment in high-risk patients; in Data Monitoring in Clinical Trials: Springer, 2006, pp 136–147.
119. Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. Taylor & Francis, 1999.
120. Proschan MA, Lan KKG, Wittes JT. Statistical Monitoring of Clinical Trials: A Unified Approach. Springer New York, 2006.
121. Alling DW. Early decision in the Wilcoxon two-sample test. *J Am Stat Assoc* 1963;58:713–720.
122. Alling DW. Closed sequential tests for binomial probabilities. *Biometrika* 1966;73–84.
123. Halperin M, Ware J. Early decision in a censored Wilcoxon two-sample test for accumulating survival data. *J Am Stat Assoc* 1974;69:414–422.
124. DeMets DL, Halperin M. Early stopping in the two-sample problem for bounded random variables. *Control Clin Trials* 1982;3:1–11.
125. Canner PL. Monitoring of the data for evidence of adverse or beneficial treatment effects. *Control Clin Trials* 1983;4:467–483.
126. Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. *Seq Anal* 1982;1:207–219.
127. Halperin M, Gordon Lan KK, Ware JH, et al. An aid to data monitoring in long-term clinical trials. *Control Clin Trials* 1982;3:311–323.
128. Lan KKG, Wittes J. The B-value: a tool for monitoring data. *Biometrics* 1988;44:579–585.
129. Cohn JN, Goldstein SO, Greenberg BH, et al. A dose-dependent increase in mortality with vesnarinone among patients with severe heart failure. *N Engl J Med* 1998;339:1810–1816.
130. Colton T. A model for selecting one of two medical treatments. *J Am Stat Assoc* 1963;58:388–400.
131. Carlin BP, Louis TA. Bayes and Empirical Bayes Methods for Data Analysis, Second Edition. Taylor & Francis, 2000.
132. Choi SC, Pepple PA. Monitoring Clinical Trials Based on Predictive Probability of Significance. *Biometrics* 1989;45:317–323.
133. Cornfield J. A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *J Am Stat Assoc* 1966;61:577–594.
134. Cornfield J. Recent methodological contributions to clinical trials. *Am J Epidemiol* 1976;104:408–421.
135. Freedman LS, Spiegelhalter DJ, Parmar MK. The what, why and how of Bayesian clinical trials monitoring. *Statist Med* 1994;13:1371–1383.
136. George SL, Li C, Berry DA, Green MR. Stopping a clinical trial early: Frequentist and bayesian approaches applied to a CALGB trial in non-small-cell lung cancer. *Statist Med* 1994;13:1313–1327.
137. Grieve AP, Choi SC, Pepple PA. Predictive Probability in Clinical Trials. *Biometrics* 1991;47:323–330.
138. Machin D. Discussion of "The what, why and how of Bayesian clinical trials monitoring". *Statist Med* 1994;13:1385–1389.
139. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: Conditional or predictive power? *Control Clin Trials* 1986;7:8–17.
140. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statist Med* 1986;5:421–433.
141. Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med* 1992;327:248–254.
142. Bauer P, Kohne K. Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics* 1994;50:1029–1041.
143. Berry DA. Adaptive clinical trials: the promise and the caution. *J Clin Oncol* 2011;29:606–609.
144. Burman CF, Sonesson C. Are Flexible Designs Sound? *Biometrics* 2006;62:664–669.

145. Chen YH, DeMets DL, Lan KK. Increasing the sample size when the unblinded interim result is promising. *Stat Med* 2004;23:1023–1038.
146. Cui L, Hun HMJ, Wang SJ. Impact of changing sample size in a group sequential clinical trial. Proceedings of the Biopharmaceutical Section, American Statistical Association, 1997 , 52–57. 1997.
147. Cui L, Hung HMJ, Wang SJ. Modification of Sample Size in Group Sequential Clinical Trials. *Biometrics* 1999;55:853–857.
148. Fisher LD. Self-designing clinical trials. *Statist Med* 1998;17:1551–1562.
149. Fleming TR. Standard versus adaptive monitoring procedures: a commentary. *Statist Med* 2006;25:3305–3312.
150. Hu F, Zhang LX, He X. Efficient randomized-adaptive designs. *Ann Stat* 2009;2543-2560.
151. Hung HMJ, Wang SJ. Sample Size Adaptation in Fixed-Dose Combination Drug Trial. *J Biopharm Stat* 2012;22:679–686.
152. Irle S, Schafer H: Interim design modifications in time-to-event studies. *J Am Stat Assoc* 2012;107:341–348.
153. Lan KKG, Trost DC. Estimation of parameters and sample size re-estimation. Proceedings–Biopharmaceutical Section American Statistical Association, 48–51. 1997. American Statistical Association.
154. Levin GP, Emerson SC, Emerson SS. Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation. *Statist Med* 2013;32:1259–1275.
155. Lui KJ Sample size determination under an exponential model in the presence of a confounder and type I censoring. *Control Clin Trials* 1992;13:446–458.
156. Luo X, Li M, Shih WJ, Ouyang P. Estimation of Treatment Effect Following a Clinical Trial with Adaptive Design. *J Biopharm Stat* 2012;22:700–718.
157. Mehta CR. Adaptive clinical trial designs with pre-specified rules for modifying the sample size: a different perspective. *Statist Med* 2013;32:1276–1279.
158. Posch M, Proschan MA. Unplanned adaptations before breaking the blind. *Statist Med* 2012;31:4146–4153.
159. Proschan MA, Hunsberger SA. Designed Extension of Studies Based on Conditional Power. *Biometrics* 1995;51:1315–1324.
160. Proschan MA, Liu Q, Hunsberger S. Practical midcourse sample size modification in clinical trials. *Control Clin Trials* 2003;24:4–15.
161. Proschan MA. Sample size re-estimation in clinical trials. *Biom J* 2009;51:348–357.
162. Shen Y, Fisher. Statistical Inference for Self-Designing Clinical Trials with a One-Sided Hypothesis. *Biometrics* 1999;55:190–197.
163. Tsiatis AA, Mehta C. On the Inefficiency of the Adaptive Design for Monitoring Clinical Trials. *Biometrika* 2003;90:367–378.
164. van der Graaf R, Roes KC, van Delden JJ. Adaptive trials in clinical research: scientific and ethical issues to consider. *JAMA* 2012;307:2379–2380.