Adaptive Methods: Telling 'The Rest of the Story'

Scott S. Emerson, MD, PhD
Thomas R. Fleming, PhD

Department of Biostatistics
University of Washington
Seattle, WA 98195-7232

Correspondences and requests for reprints should be addressed to: Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195-7232; Email: semerson@u.washington.edu

Abstract

The FDA draft guidance on adaptive design randomized clinical trials provides in depth consideration of the difficulties that unblinded adaptation of clinical trial design might introduce. We provide extended discussion of these difficulties, with focus on the problems that the adaptive designs pose in the scientific interpretation of randomized clinical trial results, for regulatory authorities as well as for patients and caregivers who wish to make evidence based decisions regarding the choice of treatment. We consider implications in adequate and well-controlled studies of the use of unblinded measures of treatment effect to make adaptive selection / modification of treatments, adaptive selection of primary endpoints, adaptive modification of maximal sample size, adaptive modification of randomization ratios, and adaptive modification of target populations (adaptive enrichment), and then we consider the special topic of seamless phase 2-3 designs. We examine the extent to which the adaptive designs do not meet the goals of having greater efficiency, being more likely to identify truly effective treatments, being more informative, and providing greater flexibility. We fully support the FDA's continued requirement of adequate and well controlled confirmatory studies, complete with prospective, detailed specification of the entire randomized clinical trial design in such a way that allows accurate and precise estimation of treatment effectiveness.

Word count: 6624 (article); 207 (abstract)

Key words: drug discovery; sequential clinical trials; exploratory; confirmatory; seamless designs; enrichment; pre-specification

1. Introduction

Randomized clinical trials (RCTs) are a mainstay in the process of evaluating new therapies prior to their adoption for use for the treatment of patients by the medical community. Owing to the ethical and economic issues inherent in these experiments conducted in human volunteers, there has long been an interest in approaches that would facilitate the more rapid, efficient completion of the therapy discovery process in a scientifically sound and statistically credible manner. Building on the sequential probability ratio test [1] and the later description of methods for computation of the exact sequential sampling distribution [2], a rich body of literature on group sequential clinical trials was developed that described families of sequential sampling plans (stopping rules), flexible implementation of those stopping rules to allow for the uncertainties in key parameters at the time of study design, and inferential methods that properly adjust for the sequential sampling density in the computation of point estimates, confidence intervals, and p values [3-4]. These methods are now widely available in statistical software [5-8].

There is also now a large statistical literature on alternatives to group sequential methods. Over a similar timeframe, a number of authors have described Bayesian methodology for the conduct of sequential clinical trials [9], and such methods have seen some use, particularly in the investigation of new devices. More recently, a number of authors proposed what have come to be known as “adaptive design clinical trials”. (See general references in [10]). These methods, both frequentist and Bayesian, have not yet seen wide acceptance, having been criticized on their ability to adequately address the scientific and statistical issues of RCTs [11-14]. Owing at least in part to perceived gaps in this literature with respect to the full complement of statistical methods needed in a regulatory setting, the U.S. Food and Drug Administration (FDA) Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER) have jointly issued a draft Guidance for Industry on “Adaptive Design Clinical Trials for Drugs and Biologics” [10].

This commentary is a review of that draft guidance. We note that the legislation regarding the FDA’s oversight of devices uses different standards and wording. Hence, although we believe the scientific and statistical principles espoused in this commentary generally apply equally well to the approval of new devices, we have directly addressed only the settings regulated by CBER and CDER.

The key points of the FDA draft guidance to our mind are its emphasis on 1) the need for unbiased and interpretable trial results from adequate and well-controlled studies that provide substantial evidence of effectiveness to support approval of a drug indication, 2) the need for prospective, detailed specification of any potential adaptation prior to any unblinding of the personnel responsible for implementing modifications of the RCT design, and 3) the potential for the counterproductive impact of adaptive design relative to the more conventional study design, owing in part to aspects of adaptive designs that are currently poorly understood. Adaptive methods seem at first blush attractive in terms of their enhanced flexibility, and the features of many such methods that would seem to allow integrating exploratory elements into what are meant to be confirmatory trials. However, there is an important need, in the words of Paul Harvey, "to tell the rest of the story" regarding the formal properties of such methods, properties that often are either inadequately understood or are unfavorable.

In order to place our comments on the draft guidance in context, we find it useful to first review in section 2 the typical process by which a new drug or biologic therapy is approved for use in the medical treatment of some disease along with a discussion of the statistical measures most often used in the regulatory process. We then in section 3 describe the general statistical approaches that are typically included as adaptive design clinical trials, briefly highlighting some of the statistical issues that the literature on those methods have and have not addressed. In section 4, we consider implications of the draft guidance with respect to the use in adequate and well-controlled studies of unblinded measures of treatment effect to make adaptive selection / modification of treatments, adaptive selection of primary endpoints, adaptive modification of maximal sample size, adaptive modification of randomization ratios, and adaptive modification of target populations (adaptive enrichment), and then we consider the special topic of seamless phase 2-3 designs. We conclude with some brief comments on the current ability of adaptive designs to meet their stated goals of flexibility and improved efficiency.

2. Drug Discovery Setting

Ultimately, a new treatment is characterized by its "indication", which will consist of 1) the exact medical definition of "disease" that might range from signs and symptoms (e.g., fever, cough) to presumed causative agents (e.g., bacterial pneumonia) that are perhaps broadly grouped by common response to a treatment (e.g., gram negative pneumonia) or narrowly focused according to lack of response to other treatments (e.g., methicillin-resistant *staphylococcus aureus*); 2) the population of patients, which might be restricted owing to concomitant medical conditions (e.g., pregnancy, renal failure) or prior failure of standard therapy (e.g., lack of tolerability); 3) the intervention itself consisting of a formulation, mode of administration, dose or dosing strategy, frequency, duration, and ancillary treatments (either prohibited due to drug-drug interactions or prescribed for prophylaxis of or rescue from adverse events); and 4) an intended outcome that materially affects the clinical manifestations of the disease.

Once a treatment indication has been approved, the actual use of the treatment will ideally be governed by "evidence based medicine". Evidence based medicine is often based on a stepwise process which closely parallels the parts of a treatment indication described above. Using the acronym "PICO", these steps consider 1) the patient (population), 2) the intervention, 3) the comparison treatment (alternatives to the intervention that might be considered), and 4) the outcome (the clinical condition that is desired). In describing the characteristics of the patient, evidence based medicine considers both the definition of the disease and any restrictions on patient characteristics apart from disease manifestations. This is in concert with the desire to refine treatments to a "personalized medicine". A clinician must not only know that a treatment has passed some standard necessary for approval of the treatment but also for each patient be able to judge the magnitude of treatment effect on outcome in a relevant patient population and to compare that expected effect to that expected for alternative treatments.

In the United States, the FDA is charged with ensuring the proper labeling, safety, and effectiveness of new drugs and biologics. The 1962 Kefauver – Harris Amendment stipulates that the no drug should be approved if "there is a lack of substantial evidence that the drug will have the effect it purports or is represented to have", where "'substantial evidence' means evidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved". More recently the FDA Amendments Act of 2007 has strengthened the authority of the FDA ensure that the benefits of drug therapy outweigh the risks in the population of patients that are likely to use the drug, including special populations and off-label use.

In order to sequentially investigate the safety and efficacy/effectiveness issues in a manner that protects the human subjects from harm, the process of investigating new treatments typically goes through a phased series of clinical trials. *Phase I* clinical trials provide safety data in support of further testing in later phases with larger sample sizes at one or more doses found to pass preliminary thresholds for incidence of adverse effects. *Phase II* clinical trials seek further systematic collection of safety data and preliminary evidence in support of biological effect. Products that fail to demonstrate a certain level of biological activity might be abandoned. Such a screening process can be shown to be more efficient in finding effective treatments from a large population of ideas.

Even when the phase II clinical trials demonstrate a desired effect on the biologic endpoint, it is common for investigators to use the results of the clinical trial to identify a more precise definition of the disease characteristics that would indicate the types of patients likely to benefit most from the treatment, a more refined definition of the population to be treated in order to eliminate subjects who might experience greater toxicity, a single treatment regimen (dose or dosing strategy, frequency, duration, ancillary prophylactic or rescue therapies), and a clinical measure to serve as the primary endpoint, as well as a statistical measure to summarize the distribution of that clinical endpoint across subjects. Such refinements of a proposed treatment indication become the pre-specified hypothesis in the large, confirmatory *Phase III* clinical trials meant to establish an acceptable benefit/safety profile in order to gain regulatory approval for a precisely defined indication (*"registrational"* clinical trials). Based on their governing laws, the FDA typically requires at least two adequate and well-controlled investigations of the effectiveness of the new drug or biologic in the setting in which the treatment will generally be used. Such independent substantiation reduces not only the possibility of false positive effects in the first trial, but also reduces the possibility that true positive results in the first study might be relevant only for the clinical sites, patient populations, investigators, and methods used in that study [15].

3. Adaptive Design Clinical Trials

In a conventional group sequential RCT design (which includes a fixed sample design as a special case), data $X_i \sim F(x;.\ \theta)$, $i=1,2,\ldots$ relevant to the estimation of some measure θ of the true treatment effect is collected and analyzed to determine whether the RCT should be terminated or whether data collection should continue to a future analysis. Ideally, θ has an interpretation useful for evidence based medicine: in the type of patient accrued to the clinical trial, a comparison of the experimental and control treatments would result in a contrast θ in the tendencies for the primary outcome. In an adequate and well-controlled study testing H_0: $\theta = \theta_0$, a number of mechanisms are instituted to ensure trial integrity and minimize the possibility that interim analyses might introduce bias into trial results. Chief among these are the pre-specification of a primary measure of treatment effect, a statistical analysis plan for estimating the treatment effect, a stopping rule for identifying the conditions under which the trial would be terminated or continued[16], and a strategy to minimize the possibility that dissemination of interim results might lead to unintended alteration of the study conduct. Up to J interim analyses will be conducted at cumulative sample sizes $N_1, \ldots, N_J$ (both J and the N_j s may be random variables so long as they are determined in a manner that is independent of the estimate of treatment effect), and the RCT can be terminated at the jth analysis or continued to the $(j+1)$th analysis depending upon the value of statistic $T_j = T(X_1, \ldots, X_{Nj})$.

The motivation for adaptive re-design of an ongoing clinical trial is varied. Upon observing interim estimates of the treatment effect that are either higher or lower than anticipated, clinical trialists might want to alter the scientific hypotheses of interest, the randomization scheme, the rules for determining the maximal statistical information to be accrued, the analysis times, the stopping boundaries, or the statistical analysis model [13, 17]. In the general adaptive setting, we thus consider $X_i \sim F_j(x;. \theta_j)$ for $N_{j-1} < i \leq N_j$, and consider the J null hypotheses H_{0j}: $\theta_j = \theta_{0j}$. It should be noted that if the full range of adaptation is considered, both the jth stage estimand (measure of treatment effect) θ_j and incremental sample size $N^*_j = N_j - N_{j-1}$ at the jth stage are actually functions of the prior data. Multiple methods of controlling experimentwise type I error for such adaptive design RCT have been proposed.

Under the J null hypotheses H_{0j}: $\theta_j = \theta_{0j}$, the unconditional distribution of incremental standardized Z statistics Z^*_j computed using only X_i, $i=N_{j-1}+1, \ldots, N_j$ is $Z^*_j \sim N(0,1)$, and the incremental fixed sample p values P^*_j are unconditionally distributed according to independent uniforms $P^*_j \sim U(0,1)$. These results serve as the foundations for the approaches to control type I error for adaptive design RCTs based on R.A. Fisher’s method for combining independent p values [18] and the “self designing” trial using weighted combinations of independent Z statistics [17].

The majority of the statistical literature on adaptive design clinical trials considers adaptive changes only to the maximal sample size, with particular emphasis on the setting in which $N^*_j = N^*_j(Z_{j-1})$, but $F_j = F$ and $\theta_j = \theta$. In that setting, it is sufficient (but not necessary) to choose second stage critical values when using a suitably chosen, pre-specified "conditional error function" [19]. A common method of implementing such an approach in adaptive design RCT with multiple stages is to pre-specify some conventional group sequential test, and then use the conditional power function from that test as the conditional error function when modifying the sample size [20].

In none of these approaches is the distribution of the test statistic known unless the entire adaptive plan is pre-specified.

4. Some Specific Topics

In section IV.A of the draft guidance, the FDA notes that the use of adaptive design in adequate and well-controlled confirmatory studies give rise to the possibility of 1) inflation of the probability of falsely declaring a treatment effective, and 2) greater difficulty interpreting study results.

With respect to the control of the type I error, we note that the proposed adaptive design methods that have received the most attention [17-20] adequately control the experimentwise type I error for the strong null hypothesis that all J of the stagewise null hypotheses H_{0j}: $\theta_j = \theta_{0j}$ are true. This in turn means that if the lowest p value were observed at the Kth stage, then the type I error is also controlled for inference about the corresponding estimand, i.e. for the weaker null hypothesis H_{0K}: $\theta_K = \theta_{0K}$. In general, however, we cannot guarantee control of the type I errors for any of the weaker null hypotheses for estimands used at the other stages, unless those estimands are identical to the Kth. In fact, it is entirely possible that adaptive design methodology based on combinations of stagewise independent standardized Z statistics or p values might indicate "efficacy" by rejecting the strong null hypothesis, yet the true value of one of the estimands estimated at a particular stage would correspond to harm.

When each stage of an adaptive design is based on a different estimand, the interpretability of the statistical inference is similarly hampered. If a treatment is to be adopted for an indication corresponding to the estimand of the Kth stage, it is not at all clear the relevance of the results from any stage which focused on some other estimand. The stage K with the lowest p value is a random variable, hence the estimand corresponding to that stage might also be a random variable. In that setting, making inferential statements about patient outcomes conditional on that indication are not easily characterized statistically, and the sampling distribution of estimated treatment effects are dependent upon the true values of the estimands from all stages prior to the *K*th. Even if only the randomization ratios at each stage are adaptively modified, the use of weighted combinations of the stagewise statistics raises the possibility of confounding treatment effects with study characteristics that vary over time, if those study characteristics are predictive of outcome. Similarly, time varying study characteristics that modify the treatment effect can lead to difficulties with interpretation even if it is only the sample sizes that are adaptively modified. These issues are of particular concern when the modifications implemented during the course of the study are highly predictive of the interim trial results: clinical investigators may modify the types of patient accrued to the study in response to their perceptions about interim results. The integrity of the A&WC study not only requires confidentiality be maintained in order to enable proper implementation of the adaptive methods (as strongly emphasized by the draft guidance in *l 1684-1699*), but also requires that the broader scientific, clinical, regulatory and investment communities do not gain indirect or direct insights about relative efficacy and relative safety of study interventions as a result of changes occurring during ongoing trials that have adaptive designs.

Below we elaborate on six specific areas where these issues might be of particular concern.

4.1 Adaptive Selection / Modification of Treatments

Adaptive design has been proposed for selection of optimal doses or other aspects of treatment delivery such as frequency, duration, or mode of administration. The issues noted above with the control of type I error and interpretability of treatment effects when the primary scientific hypotheses are adaptively modified hold in this setting. In particular, any adaptive dose selection will lead to estimates of treatment effect that have a sampling distribution dependent upon the dose-response curve at least in the neighborhood of the selected dose. Obtaining accurate and precise estimates of the treatment effect for the indicated dose is thus more difficult.

However, as noted in section VI.A. of the draft guidance, greater understanding of the effect on clinical outcomes for variation in these treatment parameters is beneficial for regulators when approving labeling guidance and for clinicians when deciding among alternative treatment strategies for a patient. Hence, RCT designs incorporating multiple dosing strategies can be an important part of an A&WC trial.

Nonetheless, a need to adaptively modify the treatment to address safety concerns or to achieve greater efficacy should carry with it a need to substantiate observed results with a later confirmatory adequate and well-controlled trial.

4.2 Adaptive Selection of Primary Endpoints

The primary endpoint in an A&WC trial evaluating effectiveness should consider those measures that best capture clinical well-being of the patient, that the treatment is likely to affect, and that can be estimated with statistical precision, in that order of importance. While there are times that multiple clinical endpoints are of nearly equal importance, it is nevertheless the case that treatments are most often ultimately administered to a patient in hopes of achieving some particular clinical outcome. Adaptive selection of endpoints therefore suffers from the same drawbacks as other adaptive modifications of the scientific hypotheses. Furthermore, because the type I error is not controlled for the stagewise estimands, statistical methods for controlling the type I error based on combinations of stagewise statistics do not in general protect against a treatment causing harm for some of the candidate estimands and benefit for others.

For instance, at an interim analysis of a phase III RCT of laromustine plus cytosine arabinoside versus cytosine arabinoside monotherapy in acute myeloid leukemia, results showed the laromustine arm with an improvement on the primary endpoint of 'objective response rate' (ORR) of 37% vs. 19% (one-sided upper $p=0.004$), but with worsening on the secondary endpoint of survival of 61% vs. 91% (one-sided upper $p > .9999$) [21]. Although such an adaptive design was not actually used in this trial, it is relevant to consider how such results might have been reported when using an approach based on R.A. Fisher's combination of p values [18]. Suppose, for instance, the first stage primary endpoint had been survival. Then at the interim analysis, the investigators might have noticed an improved ORR, and adaptively changed the endpoint for the second stage to ORR. If results for ORR at the second stage were similar to those observed at the first stage, the product of p values of $0.9999 \times 0.004 = 0.004$ is compared to the level 0.05 critical value of 0.0087, with a resulting decision to reject the null hypothesis of no beneficial effect of laromustine in favor of the one-sided alternative of a beneficial effect. It is worth noting, however, that if the desire had been to demonstrate a beneficial effect of cytosine arabinoside monotherapy, we would have combined lower p values of $0.0001 \times 0.9960 = 0.0001$, which when compared to the critical value of 0.0087 would suggest rejection of a null hypothesis of no beneficial effect of monotherapy in favor of the one-sided alternative of a beneficial effect of the monotherapy. Other methods of controlling the experimentwise type I error when using adaptive modification of primary endpoints can similarly be manipulated to obtain paradoxical results.

4.3 Adaptive Modification of Maximal Sample Size

The majority of the statistical literature on adaptive design clinical trials considers adaptive changes only to the maximal sample size. As noted in the draft guidance, sample size modification based on blinded data is well understood and is unlikely to introduce bias. The complicating factor with adaptive design trials is the use of an unblinded interim estimate of the treatment effect. Motivation for such adaptation is usually ascribed to the observation of a treatment effect that is unexpectedly less than the design alternative. Of course, such a possibility should be regularly anticipated: the trial is designed to discriminate between a null hypothesis of no effect and some design alternative of a clinically important effect. It should not therefore be surprising that the true effect might be somewhere in the middle, and proper evaluation of a clinical trial design might obviate the need for adaptive design trials [13,16].

The issue that arises with adaptive modification of maximal sample size relates less to our ability to control the type I error, as the validity (but not the efficiency) of such methods has been well-established. The larger problems relate to 1) their tendency to complicate the interpretation of trial results, 2) their tendency to publicly reveal interim results of RCT as the sample size is modified, 3) their potential to pursue effect sizes that may not be clinically relevant, and 4) their tendency toward statistically inefficient inference.

Proschan and Hunsberger [19] demonstrated that an adaptive modification of the stage 2 sample size based on stage 1 estimates of treatment effect can yield a type I error as high as 0.0616 when an unadjusted standardized *Z* statistic at the second stage is compared to *1.96,* the value associated with a nominal fixed sample type I error of 0.025. At first glance, this is quite paradoxical: The user only performed two analyses of the data, yet a Bonferroni correction based on two analyses does not protect the type I error. The paradox can perhaps be resolved, however, by considering the stochastic nature of the two analyses that were actually performed. In fact, the user considered an extremely large number of analyses (one at each possible sample size), but was able to avoid performing some of those analyses based on a prediction that significant results would probably not be attained. The ultimate type I error is seemingly affected by many more analyses than were actually performed, and the end result is an inefficient design owing to the imprecision of the “imputation” that uses early, highly variable estimates of treatment effect.

A common method of controlling the type I error without pre-specifying the exact adaptive plan is to instead pre-specify some conventional group sequential test, and then maintain the conditional power function from that test when modifying the sample size [20]. When the sample sizes are modified in this manner, the test statistics weight some observations more heavily than others, thereby violating the sufficiency principle. While such weighting controls the type I error under arbitrary adaptations, it does not lend itself to as accurate and precise point and interval estimates of treatment effect as might be possible using the sampling distribution of the sufficient statistic in a properly pre-specified adaptive design [22]. Even with the use of the sufficient statistic, the adaptive sampling plan may not be particularly efficient. Several authors have shown that a conventional group sequential design can often improve on an adaptive design trial [11, 12].

The draft guidance calls for A&WC confirmatory trials to be based on a prospectively planned adaptation with all details specified prior to unblinding. If the pre-specified design involves a large number of possible sample sizes, the sample size modification will effectively reveal the interim estimates of treatment effect. Such dissemination of interim results is widely regarded as problematic owing to the possibility that it would lead to operational biases due to changes in the behavior of investigators and/or patients with respect to participation and retention in the trial.

4.4 Adaptive Modification of Randomization Ratios

Some authors have described the use of adaptive modification of randomization ratios based on the estimates of the treatment effect on the study [23, 24]. Such an approach is not directly related to efficiency of the trial design, as the most efficient design would have samples sizes for each arm in proportion to the square root of the variability of the observations. Instead the primary goal is more directed toward individual ethics: such an approach will tend to minimize the proportion of patients on study who receive an inferior treatment. Care must be exercised however to ensure that clinical trial results are scientifically and statistically credible to a broad audience, and that the adaptation of the randomization ratio does not introduce operational bias into the trial. These issues are illustrated in the controversies surrounding the evaluation of extra-corporeal membrane oxygenation (ECMO) in infants [25]. A randomized play the winner adaptive design conducted in 12 infants resulted in a single patient receiving usual standard of care (SOC) and 11 infants receiving ECMO, with the SOC patient dying and all ECMO patients surviving. Despite the prior belief that the patient population would experience >80% mortality under SOC, scientists could not agree on the effectiveness of ECMO in part because of disagreement on a p value to ascribe to these results. Suggested values ranged from p=0.001 to p=0.62 [26]. Ultimately, a second RCT was performed in which 6 of 10 patients treated with SOC survived, while all 9 patients treated with ECMO survived [25]. Then a much larger RCT in the U.K. showed 38 of 92 patients on SOC survived and 65 or 93 patients on ECMO survived [27].

It is not at all clear that the initial randomized play the winner design contributed substantially to the ethical treatments of infants with respiratory failure from a population ethics standpoint. To the extent that we could be sure that the underlying mortality rate under SOC in those first randomized patients was 80%, then the vast majority of patients on the trial would seem to have benefitted. However, the study results were judged inconclusive in part because one could not be certain of that baseline rate. As the trial progressed, the type of patient accrued to the study might have changed. Certainly it would appear that the patients accrued to the later trials did not experience 80% mortality under SOC. Hence, many would argue that skepticism of the initial trial results was warranted. It remains an open question as to whether a more conventional RCT conducted from the very first would have saved some part of the 10 years that it took to mount the confirmatory trials.

This trial experience also highlights the importance of considering time trends when using adaptive design clinical trials. If the initial estimates of 80% mortality on SOC are valid, then there was certainly a time trend in survivability on SOC among patients accrued to RCT over the course of the three clinical trials. To the extent that any part of that increase in survivability occurred within a single study, the changing randomization ratio means that the treatment was confounded with patient characteristics related to the primary outcome. In order to avoid this potential bias, analysis would need to stratify subjects within groups sharing common randomization ratios, thus complicating the interpretation of observed treatment effects.

4.5 Adaptive Modification of Target Population (Adaptive Enrichment)

With increased interest in targeted therapies, there is recognition of the need for effective approaches to identify ‘enriched’ subgroups of patients who are more likely to benefit from such therapies. This is particularly important when treatments are toxic, inconvenient, or costly, and where it is likely that there are strong effect modifiers regarding efficacy. While adaptive methods have been proposed for the pursuit of enrichment, a RCT using an adaptive selection of the target population cannot be regarded as confirmatory of a pre-trial specified treatment indication.

Post-hoc exploratory subgroup analyses are treacherous due to the great risk that there will be ‘random high’ overestimates of treatment effect in subgroups having the most favorable estimates of treatment benefit, and because the absence of a proper sampling context renders the p-values for treatment effect in those subgroups uninterpretable. While pre-specification of a planned alpha-sharing between the overall analysis and an analysis in the biomarker subgroup of principal interest would address concerns about inflation of the type I error, one should carefully consider the consequences of such an approach on the ability of regulators (and, if the treatment is eventually approved, clinicians) to judge the magnitude of effect of the treatment in the eventual population. In the presence of adaptive selection of the patient population, the sampling distribution for estimates of the treatment effect in any restricted population ultimately depends on the treatment effect in each of the subpopulations considered in the adaptation.

Consider the setting where an EGFR-inhibiting drug is to be evaluated in colorectal cancer patients who have failed several prior regimens, and where interest is in effect both in the overall population as well as in the subgroup of patients whose tumors express the wild type version of the KRAS gene. Suppose an alpha-sharing approach is pre-specified where the wild type subgroup is assessed at the one-sided 0.005 level and the pooled analysis is conducted at the one-sided 0.02 level. If the data do reveal much higher estimated effects in patients with the wild type rather than mutant version of the KRAS gene, this alpha-sharing approach does provide enhanced sensitivity to these benefits in wild type patients. However, there is a price for obtaining this enhanced sensitivity in this manner. Regulatory approval likely would not be granted in the ‘mutant’ subgroup even when the pooled analysis meets the targeted one-sided 0.02 level of significance if the effects in the ‘mutant’ subgroup are sufficiently modest that it is apparent the positivity in the pooled group is driven by the favorable result in the ‘wild type’ subgroup. In such a setting, it would not be persuasive to suggest the favorable estimate in the pooled sample should apply to all patients based on the argument that the evidence of effect modification, and specifically an unimpressive estimated effect in ‘mutant’ patients, is spurious.

In settings where a monotonic relationship is expected between efficacy and the level of genetic expression, statistical methods have been proposed to enable a single trial to be used to perform both discovery (i.e., determination of the sensitive subgroup) and validation of meaningful benefit in that ‘discovered’ subgroup [28,29]. In the illustration in Figure 1, suppose genetic expression and hence level of efficacy monotonically increases as one moves from left to right. These authors have shown in extreme settings (e.g., when 90% of patients receive no benefit, i.e., $RR_l = 1$, and 10% experience a multi-fold increase in time to event, i.e., $RR_u = 0.21$) that discovery and confirmation can be effectively performed in a single trial. However, in much more likely scenarios that are shown in Figure 1, the operating characteristics of this approach are not favorable. In essence, to obtain reliable results when carrying out discovery and confirmation in a single trial, very large clinical trials would be required.

4.6 Seamless Phase 2-3 Designs

Several authors have proposed that increased flexibility and efficiency in clinical development could be achieved through “seamless Phase 2-3 designs” [30-34] that allow the cohort of Phase 2 patients to be included in the Phase 3 trial. They suggest efficiency is further enhanced because the combination of these two phases into a single trial allows the elimination of the calendar time (sometimes referred to as the “white space”) that can be lengthy between the completing of Phase 2 and the initiation of the enrollment into Phase 3.

Significant concerns arise with such designs; these concerns result from the fact that access to the results from the Phase 2 trial would be restricted to the Data Monitoring Committee (DMC), see Figure 2. The need for such a restriction arises because, in essence, analysis of the Phase 2 trial data effectively becomes an interim analysis of the Phase 3 trial, and compelling arguments exist that such interim data should remain confidential [35, 36].

Phase 2 trials not only provide insights into whether to conduct the Phase 3 trial, but also how best to design it. These insights include enhanced understanding about dose and schedule of the intervention, formulation of supportive care, types of safety concerns that may emerge and how best to detect them, defining proper primary and secondary endpoints and the proper duration and frequency of follow-up, and approaches to enhance quality of trial conduct, including improving enrollment, event rates, adherence and retention. By restricting access of these insights to the DMC upon completion of Phase 2, important responsibilities in the clinical development process regarding optimizing the design of the Phase 3 trial are inappropriately transferred from the study research team, sponsor and regulatory authorities to the DMC. It has been proposed that some of these responsibilities might be retained by the research team and sponsor through a detailed formulation of decision rules and by sophisticated statistical procedures that adjust for these decision rules in a manner to prevent inflation of false positive and false negative error rates and 'random high' bias in estimates of key parameters. Unfortunately, such decision rules cannot be sufficiently comprehensive to anticipate many areas where significant insights will emerge from a typical Phase 2 trial, resulting not only in exclusion of the sponsor from this critically important decision making step in the clinical development process but also in sub-optimal ability to refine the Phase 3 trial based on the broader insights provided by the Phase 2.

Even the adaptive methods that include pre-specified detailed decision rules and sophisticated statistical procedures to provide adjustment for these decision rules may, in real world settings, lead to meaningful reductions in the integrity and interpretability of results. As recognized by FDA in their draft guidance [10], the implementation of such adaptive method would need to follow the 'letter of the law' to enable FDA and others to conduct proper inferences. However, in the real world, there are irregularities in quality of trial conduct, and unexpected important insights about efficacy and safety, often in unanticipated measures, that could be meaningfully influential and could lead to important deviations from the pre-specified adaptive procedures. Because such complexities were recognized long ago, it was understood that classical group sequential monitoring boundaries could not be rigidly implemented, but rather they would serve as guidelines for DMCs that would need to be in place to enable properly informed and independent judgment regarding formulation of decisions about the 'adaptive' issue of when to terminate a trial. While this flexibility generally does not compromise the interpretability of a conventionally designed trial using a group sequential guideline, it is not at all clear that it would not meaningfully compromise the interpretability of a trial implementing more recently proposed adaptive methods.

Finally, the benefits of a seamless Phase 2-3 design in eliminating the “white space” between Phase 2 and Phase 3 trials are partially or fully lost by the substantial increase in “white space” that they induce after Phase 1. This increase post Phase 1 arises because there are much greater scientific and regulatory complexities when, based on insights from only the Phase 1 trial, one is designing and seeking approval to launch what in essence is a Phase 3 trial rather than a traditional Phase 2 study that would have allowed full access to data upon its completion.

5. Discussion: Do Adaptive Designs Meet Their Goals?

The draft guidance on adaptive design RCT presents an admirable summary of the difficulties that unblinded adaptation of clinical trial design might introduce into A&WC RCTs, as well as the roles that adaptive design might have in the earlier, exploratory RCT. In our commentary we have stressed the problems that the adaptive designs pose in the scientific interpretation of RCT results, both from a regulatory perspective and from the perspective of the information needed to evaluate the best approved therapy to use in treating a patient. We are fully supportive of the FDA’s continued requirement of A&WC confirmatory studies, complete with prospective, detailed specification of the entire RCT design in such a way that allows accurate and precise estimation of treatment effectiveness.

It is also of interest, however, to examine the extent to which the adaptive designs do not meet the goals of having greater efficiency, being more likely to identify truly effective treatments, being more informative, and providing greater flexibility. In the previous section, we have identified several aspects of adaptive designs that would be expected to lead to less efficiency and power than other more conventional designs. We have previously characterized this among the “costs of planning not to plan” [13]. We have also identified many aspects of adaptive designs that we believe contribute less, rather than more, of the information needed to evaluate the effectiveness of studies. We believe that much of this diminished information is due to the misguided overemphasis on achieving statistical significance rather than on achieving statistically reliable evidence of clinically meaningful effects. To be specific, the goal of clinical research is not simply to achieve statistical significance; rather "*The primary goal should be to obtain a statistically reliable evaluation regarding whether the experimental intervention is safe and provides clinically meaningful benefit*" [14]. Lastly, we feel that there is a real possibility that the use of adaptive designs removes some of the flexibility that a RCT needs to react to external information. As described in the draft guidance, RCT design modification that is motivated by changes in the clinical setting independent of the RCT pose no great problem with respect to bias or statistical inference. However, to the extent that that external information is also related to aspects of the RCT that might be adaptively modified using unblinded trial data, the ability to incorporate that external information in an efficient manner can be greatly hampered.

Adaptive design RCTs may indeed have an important place in the early, exploratory studies used to generate hypotheses to be tested in an A&WC confirmatory trial. However, even then, we find it useful to note that the major determinant for obtaining statistically significant results in a confirmatory trial is to have a truly effective treatment. And the proper use of early screening trials to "weed out" ineffective therapies is our major tool in increasing the prevalence of truly effective therapies among those that advance to phase III trials. Hence, when many authors decry the low rate of "positive" phase III studies and use that low rate as justification for increasingly "innovative" adaptive procedures, we believe they may be placing their efforts in the wrong place. The goal of A&WC RCT should be to ensure that nearly all "positive" studies are in fact true positives. This is best achieved when we use our early phase studies in a systematic fashion to enrich the prevalence of truly effective treatments studied in phase III, and study designs that increase the type I error and type II error of those early phase studies will lead to fewer "positive" A&WC trials.

Thomas Edison once said, "Opportunity is missed by most people because it is dressed in overalls and looks like work." In clinical science, it is the steady, incremental steps that are likely to have the greatest impact.

6. References

1. Wald A. *Sequential Analysis.* Wiley, 1947.

2. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A*, **132**: 235-244, 1969.

3. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. CRC Press 2001.

4. Whitehead J. *The Design and Analysis of Sequential Clinical Trials.* John Wiley & Sons, 1997.

5. PEST (Planning and Evaluation of Sequential Trials).The MPS Research Unit, Lancaster University, Lancaster, U.K., 2008

6. East. The Cytel Software Corp. Cambridge, Massachusetts, 2010.

7. S+SeqTrial. (2002) TIBCO, Seattle, Washington, 2002

8. SAS PROC SEQDESIGN, PROC SEQTEST, SAS Institute, Cary, North Carolina 2009

9. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaces to Clinical Trials and Health-Care Evaluation.* Wiley, 2004.

10. Draft FDA Guidance Document: Guidance for Industry Adaptive Design Clinical Trials for Drugs and Biologics, Feb. 2010, http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

11. Tsiatis AA, Mehta CR. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, **90**:367–378, 2003.

12. Jennison C, Turnbull BW. Adaptive and non-adaptive group sequential tests. Biometrika 93, 1-21, 2006.

13. Emerson SS. Issues in the use of adaptive clinical trial designs. *Statistics in Medicine* 25: 3270-3296, 2006.

14. Fleming TR. Standard vs. adaptive monitoring procedures: a commentary. *Statistics in Medicine*. 25: 3305-3312, 2006.

15. FDA Guidance Document: Guidance for Industry Providing Clinical Evidence of Effectiveness for Human Drugs and Biological Products, May, 1998 http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

16. Emerson SS, Kittelson JM, Gillen DL. Frequentist evaluation of group sequential designs. *Statistics in Medicine,* 26(28): 5047-80, 2007.

17. Fisher LD. Self-designing clinical trials. *Statistics in Medicine*, **17**: 1551–1562, 1998.

18. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**: 1029-1041, 1994.

19. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics*, **51**: 1315–1324, 1995.

20. Müller H, Schäfer H. Adaptive group sequential designs for clinical trials: combining the 1968 advantages of adaptive and of classical group sequential approaches. *Biometrics* 57, 886-891, 2001.

21. DeAngelo D, O'Brien SM, Vey N, Seiter K, Stock W, Cahill A, Pigneux A, Claxton D, Stuart R, Giles FJ. A double blind placebo-controlled randomized phase III study of high dose continuous infusion cytosine arabinoside with or without VNP40101M in patients with first relapse of acute myeloid leukemia..*Journal of Clinical Oncology,* **26** (May 20 suppl; abstract 7051), 2008.

22. Shi S. Estimation following self-designing clinical trials. Unpublished{MS Thesis, Department of Biostatistics, University of Washington, Seattle, Washington, 2003

23. Berry DA, Eick SG. Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine* 14: 231-246, 1994.

24. Yao Q, Wei LJ. Play the winner for phase II/III clinical trials. *Statistics in Medicine* 15: 2413-2423, 1996.

25. Ware JH. Investigating therapies of potentially great benefit: ECMO. *Statistical Science* **4**:306-316, 1989.

26. Begg CB. Comment on Investigating therapies of potentially great benefit: ECMO. *Statistical Science* **4**:320-322, 1989

27. UK Collaborative ECMO Trial Group. UK collaborative randomised trial of neonatal extracorporeal membrane oxygenation. *Lancet* **348**:75-82, 1996.

28. Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Instit* 99:1036-1043, 2007.

29. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 11: 7872-7878, 2005.

30. Inoue LY, Thall PF, Berry DA. Seamlessly expanding a randomized phase II trial to phase III. *Biometrics* 58: 823-831, 2002.

31. Berry DA. Bayesian Statistics and the Efficiency and Ethics of Clinical Trials *Statistical Science*, Vol. 19, No. 1 (Feb., 2004), pp. 175-187. Published by: Institute of Mathematical Statistics

32. Berry, DA et al. in Case Studies in Bayesian Statistics V, (eds Gatsonis, Cl, Carlin, B. & Carriquiry, A.) 99-191 (Springer, New York, 2001).

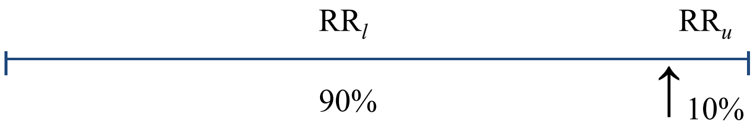
33. Bretz F, Schmidli H, Konig F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypothesis selection at interim: general concepts. *Biometrical Journal* 48(4): 623-634, 2006.

34. Jennison C, Turnbull BW. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: opportunities and limitations. *Biometrical Journal* 48: 650-655, 2006.

35. Ellenberg SS, Fleming TR, DeMets DL: Data Monitoring Committees in Clinical Trials: A Practical Perspective. New York: John Wiley and Sons, 2002.

36. Fleming TR, Sharples K, McCall J, Moore A, Rodgers A, Stewart R. Maintaining Confidentiality of Interim Data to Enhance Trial Integrity and Credibility. *Clinical Trials* 5: 157-167, 2008.

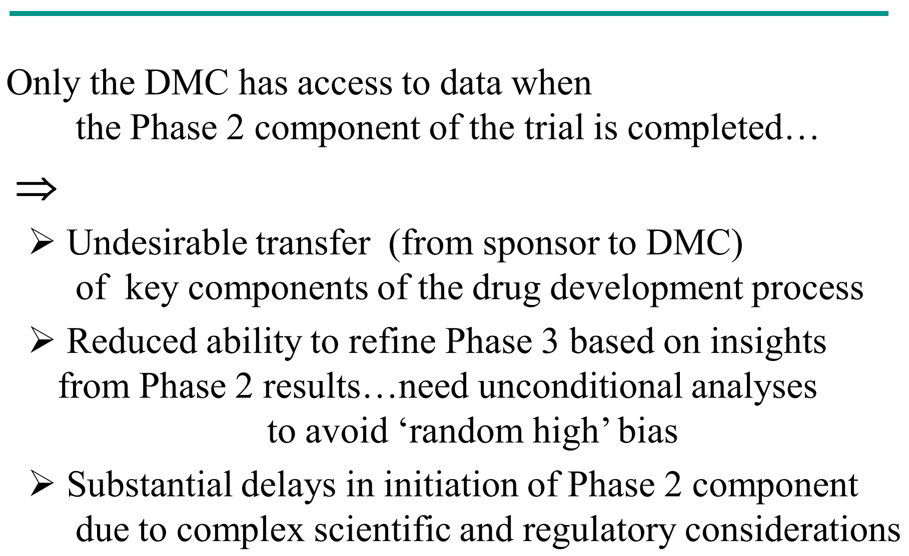
Figure 1: Operating characteristics of adaptive methods for enrichment should be considered over the wide range of scenarios that arise in real world settings. Properties are noted when using a ‘biomarker adaptive threshold’ design [28]. RR_l and RR_u denote the true relative risks in cohorts with lower and higher biomarker values, respectively.



Concerns:

- If $\mathrm{RR}_l = 1.0$ & $\mathrm{RR}_u = 0.5$ → power is low… thus, if interaction likely, restrict enrollment to targeted cohort
- If $\mathrm{RR}_l = 0.8$ & $\mathrm{RR}_u = 0.6$ → power gain is negligible
- If $\mathrm{RR}_l = 0.75$ & $\mathrm{RR}_u = 0.7$,
  → increased false negative risk in lower group
- If $\mathrm{RR}_l = 0.7$ & $\mathrm{RR}_u = 0.7$ → some loss of power
- Method → Estimates will have “random high” bias
- Bottom line: “No free lunch” from adaptive methods

Figure 2: Important negative consequences arise from the implementation of Seamless Phase 2-3 designs in clinical trials.



Only the DMC has access to data when the Phase 2 component of the trial is completed…

⇒

- Undesirable transfer (from sponsor to DMC) of key components of the drug development process
- Reduced ability to refine Phase 3 based on insights from Phase 2 results…need unconditional analyses to avoid ‘random high’ bias
- Substantial delays in initiation of Phase 2 component due to complex scientific and regulatory considerations