Published in final edited form as:

Clin Res Regul Aff. 2015 ; 32(1): 36–44. doi:10.3109/10601333.2014.977490.

Adaptive designs for comparative effectiveness research trials

John A. Kairalla¹, Christopher S. Coffey², Mitchell A. Thomann², Ronald I. Shorr^{3,4}, and Keith E. Muller⁵

¹Department of Biostatistics, University of Florida, Gainesville, FL, USA

²Department of Biostatistics, University of Iowa, Iowa City, IA, USA

³Department of Epidemiology, University of Florida, Gainesville, FL, USA

⁴Geriatric Research Education and Clinical Center (GRECC), Malcom Randall Veterans Affairs Medical Center, Gainesville, FL, USA

⁵Department of Health Outcomes and Policy, University of Florida, Gainesville, FL, USA

Abstract

Context—Medical and health policy decision makers require improved design and analysis methods for comparative effectiveness research (CER) trials. In CER trials, there may be limited information to guide initial design choices. In general settings, adaptive designs (ADs) have effectively overcome limits on initial information. However, CER trials have fundamental differences from standard clinical trials including population heterogeneity and a vaguer concept of a “minimum clinically meaningful difference”.

Objective—To explore the use of a particular form of ADs for comparing treatments within the CER trial context.

Methods—We review the current state of clinical CER, identify areas of CER as particularly strong candidates for application of novel ADs, and illustrate potential usefulness of the designs and methods for two group comparisons.

Results—ADs can stabilize power. The designs ensure adequate power for true effects are at least at clinically significant preplanned effect size, or when variability is larger than expected. The designs allow for sample size savings when the true effect is larger or when variability is smaller than planned.

Conclusion—ADs in CER have great potential to allow trials to successfully and efficiently make important comparisons.

1. Introduction and motivation

1.1 Introduction

Randomized clinical trials (RCTs) are considered to be among the most powerful and reliable tools of medical research. Important clinical trial results can have widespread

influence on clinical and health policy decisions. Traditional clinical trial methodology is designed to minimize bias and allow for strong comparison of hypothesized causal relationships. Despite their strengths, traditional clinical trial designs have a number of drawbacks in modern clinical research settings.

For one, traditional trials are not conducted in “real world” settings. RCTs depend on defined *efficacy* endpoints with varying degrees of inclusiveness. Some examine whether the treatment works under ideal, highly controlled settings, and typically use homogenous populations carefully defined by extensive and detailed inclusion and exclusion criteria. As a result, it can be difficult to determine the external validity of trial results in conditions and populations that differ from those included in the study.

Additionally, primary controlled clinical trials are usually conducted with an experimental treatment compared to a placebo, with both perhaps being supplementary to a baseline of care. Thus, the goal is to determine whether a new treatment has an incremental improvement in health outcomes versus a standard of care. However, many such potential treatments often exist and confusion arises as to which treatment is in fact best for a patient or in a population.

Another drawback for traditional designs is that the designs are rigid and success is largely dependent on a priori knowledge that is largely unknowable. In traditional clinical trials, key design elements (e.g. primary endpoint, clinically meaningful treatment difference, measure of variability, or control event rate) are pre-specified during study planning. Once all data is collected, a final analysis is performed. Consequently, study success strongly depends on the accuracy of the original assumptions. Combined with the fact that clinical trials are extremely costly and time consuming, this is a significant weakness when initial assumptions are wrong.

Two areas of research designed to address these drawbacks of traditional trial design are the fields of comparative effectiveness research (CER) and adaptive designs (ADs). This paper will examine two stage adaptive comparative effectiveness settings with continuous outcomes for comparing two active treatments. Section 1.2 introduces CER and section 1.3 introduces ADs. Following the introductions, section 1.4 describes a motivating example that introduce issues to be addressed in order achieve the benefits of CER trials. Current and ongoing research seeks to incorporate these methods into potential CER trial designs in order to efficiently allow for sample size and study duration flexibility. Section 2 describes a potential incorporation of adaptive designs into CER. Section 3 discusses evaluation and includes a selection of results showing their promise.

1.2 Comparative effectiveness research

The field of comparative effectiveness research (CER) has grown as a response to the costs and drawbacks of traditional research designs (1). The overall desire is to assist doctors and policy makers in deciding which treatments are preferred for a particular patient in a given context. Decision making in CER is typically based on head to head comparison of active treatments and the use of real-world population samples (2). The potential evidentiary and economic benefits of CER have brought it to the forefront of current medical research and it

has become a scientifically, culturally, and economically demanded part of healthcare reform. The American Recovery and Reinvestment Act of 2009 included a \$1.1 billion investment in CER and the Patient Protection and Affordable Care Act of 2010 authorized the creation of the Patient-Centered Outcomes Research Institute (PCORI) (3,4). PCORI, an independent, nonprofit health research organization, has a mission to “fund research that offers patients and caregivers the information they need to make important healthcare decisions” with a focus on “comparative clinical effectiveness research” (5). PCORI is expected to receive \$3.5 billion in appropriations funding for 2010–2019, firmly entrenching CER as a national research priority in the upcoming decade.

An Institute of Medicine Committee on Initial Priorities for Comparative Effectiveness Research (1) created a working definition of CER as:

“...the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat and monitor a clinical condition, or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels”.

We focus our attention here on the “generation and synthesis of evidence” comparing “alternative methods” allowing for “informed decisions”. Methods for reliable evidence on the comparative effectiveness of treatments for medical and health policy decision makers were deemed inadequate (6). Tunis et al. (7) provided an excellent summary of the current policy context, need for future methods development in CER, and summary of existing research infrastructure. They also clarified that CER covers a range of different approaches including: 1) systematic review, 2) decision modeling, 3) retrospective analysis of existing clinical or administrative data, 4) prospective observational studies, and 5) experimental studies, including RCTs. Although all of these areas are interesting, the focus of this manuscript lies with the use of RCTs for conducting CER. RCTs must have a prominent place in CER due to their reliable information and well respected standard. However, improved approaches and CER-focused rethinking are needed to ensure their feasibility and overcome tendencies to be slow, expensive, and homogenous in sample (6,8).

Progress towards this goal has been made with the advancement of *pragmatic randomized controlled trials* (pRCTs) for use in CER (9). Chalkidou et al. defined pRCTs as “trials that maintain the internal validity of RCTs while being designed and implemented in ways that better address the demand for evidence about real-world risks and benefits among a broad spectrum of patients, so that they will better inform clinical and health policy decisions” (9). The goal is to assess effectiveness in the “real world” outside of the highly controlled environment of modern efficacy trials. To formalize the terminology, design, and reporting of pragmatic trials, the Consolidated Standards for Reporting Trials (CONSORT) statement was extended for pragmatic trials (10) and Thorpe et al. developed the pragmatic-explanatory continuum indicator summary (PRECIS) (11). The PRECIS tool allows for graphical display of the degree of pragmatism within multiple dimensions of a trial while the CONSORT statement assists with proper reporting of results. The pRCT is designed to balance internal validity (through randomization) with external validity through generalization to real-world settings.

While the pragmatic trial approach makes progress towards that generalizability desired in CER, it does not directly address the concern that CER trials may be larger and more costly than standard trials. Potentially increased variability from heterogeneous populations, and decreased effect sizes from active comparators could mean that traditionally designed studies require more time and increased sample size. Consider a real world CER trial example, the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) (12,13). The ALLHAT study was a randomized, double-blind trial comparing traditional anti-hypertensives medicines (diuretics) and three newer medication classes. The National Heart, Lung, and Blood Institute (NHLBI) funded trial originally enrolled over 42,000 participants in order to make groupwise comparisons finding that the diuretics are at least as effective, and less costly compared to the newer medications. Despite the large size, eight years of study monitoring, and many millions of dollars spent, the study results have been questioned and had a smaller than expected clinical impact on practice (14).

Maintaining study efficiency while embracing the pragmatic nature of CER is an important statistical undertaking. As Morton and Ellenberg stated, “CER will require a proactive insertion of statistical methods that are innovative, scientifically rigorous, relevant, and timely” (15).

1.3 Adaptive designs

Adaptive designs (AD) continue to attract substantial interest in regulatory health science, as evidenced by the U.S. Food and Drug Administration (FDA) draft guidance document (16). ADs give one way to address the uncertain choices that must be made during planning for a CER trial. ADs are “adaptive” in that they allow changing characteristics of a study based on information accumulated during study implementation. Among other items, these characteristics could include study duration, sample size, or the number of study arms. ADs are “designs” in that the adaptations are planned. Consistent with FDA guidance, a PhRMA working group (17) stated that adaptive designs “...modify aspects of the study as it continues, without undermining the validity and integrity of the trial.” Additionally, they stated that “...changes are made by design, and not on an ad hoc basis”. Thus, ADs allow for *planned* modifications. The flexibility they allow can translate into a more efficient research pathway by increasing the chance of trials correctly answering questions of interest. More information on adaptive designs in general can be found in recent reviews (18,19). A wide variety of adaptive design methods could have potential useful application in CER trials. Various stage-wise p-value and test statistic combination methods are available that can control type I and type II error rates and are optimized under various conditions (20–26). Additionally, Bayesian methods for adaptive randomization or incorporating prior information into the current study context have been singled out for potential CER application (2,27,28) However, of particular interest for our current exploration of use in two-group CER trials are group sequential (GS) and sample size re-estimation (SSR) methods. Both GS and SSR were developed to create more efficiency in studies and can be seen as addressing parameter mis-specification. Each will be briefly described immediately below.

GS designs allow stopping a trial early through interim testing if it becomes clear that a treatment is superior or inferior. Thus, GS methods protect against effect size misspecification. In studies with effect size uncertainty, GS designs have proven to be an efficient method at achieving power over a range of effects while having beneficial overall power properties (29,30). Jennison and Turnbull concluded that "...when there is a choice between a minimal clinically significant effect size and larger effect sizes that investigators hope to see is that the power requirement should be set at the minimal clinically significant effect...bearing in mind that a good sequential design will reduce the observed sample size if the effect size is much larger than the minimal effect size." (30). Several approaches have been proposed to allow for repeated interim testing while preserving the type I error rate. Well known among them are the approaches described by Pocock (31), by O'Brien and Fleming (32), as well as the triangular method and double triangular method described by Whitehead and Stratton (33, 34). The method described by Pocock takes the approach of finding a single adjusted nominal significance level that can be used at each testing time. Alternatively, the O'Brien and Fleming (OF) method has the nominal significance level increase as more information accrues during the study. The OF approach has become much more popular due to the preferred characteristic of preserving power to later in a study when more information is at hand. The other boundary shape method common in practice is the triangular method (33, 34). Here straight line boundaries on the score scale are used to allow for two-sided stopping along with stopping for futility. Additionally, α -spending functions (35) are important tools that allow flexible timing of analyses regardless of the type I error preservation method employed. For two group CER trials, two-sided GS methods make the most sense, allowing for efficacy stopping for either treatment group being considered. With futility stopping included, GS methods could also then be seen as protecting against variance misspecification by allowing for early stopping if variance estimates make future stopping unlikely. A common futility method in group sequential methods that seems promising in CER is the conditional power approach to stochastic curtailment (36). GS methods in general are well known and have been extensively described (36).

SSR methods allow design parameters to be changed or re-estimated, with the study sample size adjusted accordingly. The sample size change could be based on updated values for the effect size or for other nuisance parameters (such as variability). Sample size re-estimation based on observed effect size has generated considerable discussion and controversy (29,30,37) relating both to inefficiencies and the potential conveyance of considerable information from the interim decisions that are made. However, there is little controversy concerning SSR based solely on updated nuisance parameters. These designs, known as internal pilots (IP; (38)), are two stage designs without interim testing, but with SSR based on first stage nuisance parameter estimates. The designs protect study power at the preplanned clinical effect of interest. Also, an IP implementation (with SSR based solely on the nuisance parameter) where non-objective parties have no access to raw data gives no information about effect trends of interest (18).

Recent efforts have focused on combining GS and IP based SSR designs in order to achieve their benefits simultaneously. Here, at each interim stage, stopping is allowed based on prespecified rules, and also maximum sample size for the study can be adjusted based on nuisance parameter updates from interim information. Asymptotically correct methods for

GS and IP methods in clinical trials have been proposed (39,40). The methods, however, are most effective in large trials and can have type I error rate inflation in small sample studies. Distributional theory for internal pilots with interim analysis (IPIA) allows for direct calculation of achieved power, expected sample size and error rates and has computational time advantages over simulation methods and the method has power and expected sample size benefits over fixed sample methods (41). Kairalla, Coffey, and Muller (42) identified three potential sources of type I error rate inflation in IPIA designs and showed how they could be controlled effectively. The risks include type I error rate inflation due to multiple testing, the use of asymptotic theory, and the use of downwardly biased variance estimates inherent to IP based SSR designs (42, 43).

1.4 Motivating example

As a hypothetical, motivating example, consider the planning for a two arm randomized controlled trial comparing active depression treatments with similar safety and availability (adapted from (44)). The primary endpoint for the hypothesized study is 6 week decline in the Hamilton Depression Index (HAM-D) (45). While sometimes HAM-D scores are square root transformed to induce normality, for didactic purposes, we consider the 6 week decline scores to be normally distributed without further justification. As both treatments being considered are proven active treatments, there is general uncertainty about the expected absolute group-wise mean difference ($\delta = |\mu_1 - \mu_2|$). On one hand, the investigators believe that $\delta = 1.5$ is a reasonable expectation, but they also recognize that any true value above $\delta = 1$ would be clinically non-equivalent and meaningful. That is, if proven, the knowledge could potentially change first line clinical practice recommendations. In this case, the question arises as to how to determine the sample size for the study. For example, should the reasonable expectation ($\delta = 1.5$) be used to plan a smaller study with good chance of success for higher effect sizes, but with an increased risk of missing smaller (but clinically significant) effect sizes? Or should a much larger study be planned, assuming $\delta = 1$, in order to ensure that success is achieved for smaller values, but risk being overpowered for larger effect sizes? The latter case would call for over twice the sample size commitment, using much more resources and time. In our experience, for practical reasons the larger but reasonable effect size expectation ($\delta = 1.5$) would be the most common choice made during planning. This would be especially true in head to head CER trials, where the minimal effect of interest will typically be smaller than in standard trials and may often be lower than what is deemed a reasonable effect.

2. Adaptive comparative effectiveness research trials

2.1 Some issues with comparative effectiveness research trials

The concept of a “minimum clinically meaningful difference” has diminished meaning in CER trials. Even assuming equal cost and safety, a range of meaningful effect sizes could be defined with upper limit as the largest effect with a reasonable chance of being observed and lower limit as the minimal effect deemed sizable enough to change practice in the study context. Traditional designs would have good power to detect only large effect sizes. Thus, clinically small, but population important, differences may be missed in CER trials. Finally, prior variability estimates could only be available from highly controlled studies from

homogeneous populations. Higher variability in outcomes can be expected from the “real-world” populations studied in CER. If the variance, σ^2 , is underestimated during planning, the result is an underpowered study. Alternatively, an overestimate contributes to study inefficiency through increased sample size.

Consider possible research outcomes under the HAM-D example in subsection 1.4 designed for power to detect the larger effect ($\delta = 1.5$) for practical reasons (assuming variance estimates are near pre-planned values). The best case would be obtaining a result with one treatment clearly better than the other ($\hat{\delta} = 1.5$). Here the design was appropriate and the study gives a large significant difference and a definitive answer to the research question. Another potential outcome would be finding very little difference between the two treatments. Here, the treatment effect may show a more or less equivalent effect on depression decline over the six week period. Although the investigators may be disappointed by an initial preference not being superior, they can take comfort that they have made a useful contribution to the body of knowledge comparing the two treatments and finding little difference. A more disappointing result in this setting would be one with a moderate observed treatment difference (e.g., $\hat{\delta} = 1.25$), with little chance of proving superiority. Here, there is no statistically significant difference, but there could be an unproven clinically meaningful difference present. We refer to this area as the “statistical gray zone”. As there is no default treatment choice in a CER trial, not much can be taken from this underpowered, ambiguous result other than that more research is needed for a clearer answer to the research question. Care must be taken to avoid this poor alignment between study goals and design. As mentioned above, additional uncertainties in nuisance parameter values (such as variance) could be more prevalent in CER trials. If left unaddressed, this uncertainty would only add to the misalignment of study goals and results.

Adaptive designs (ADs) have been proposed to improve design characteristics in traditional trials by allowing for greater flexibility to adjust a study based on accumulating information. Specifically, sample size re-estimation (SSR) and group sequential (GS) methods seem to hold promise in CER. The examination of the use of these ADs in CER could improve trial accuracy and efficiency in this important field.

2.2 A potential adaptive CER trial

A proposed new aspect of an adaptive CER trial would be the introduction of primary *and* secondary effects of interest. These could represent, for example, the endpoints of a range identified at the upper end by the largest reasonable expected effect and on the lower end by the smallest effect deemed sizable enough to show superiority and change primary clinical practice. Thus, an example design for two group CER trials could have two stages with efficient group sequential design based stopping bounds and total sample size to control the type I error rate and allow for two testing stages. The first stage sample size is chosen to detect the larger reasonable effect size (such as 1.5 points in HAM-D reduction difference) and the second sample size is chosen so that overall power is controlled at the smaller secondary effect. At the conclusion of the first stage, researchers could 1) declare efficacy (one treatment clearly better), 2) declare futility (study unlikely to show difference between treatments), or 3) proceed with a second stage powered to detect the secondary effect if

results suggest a smaller, but meaningful, effect might exist. Thus, the range of effect sizes of interest is covered by the study design, with smaller studies more probable if effect sizes are large. If variance is also uncertain, additional SSR calculations for the second stage sample size could also incorporate observed information based on the first stage data for adaptively determining the second stage sample size. The larger the first stage sample size, the greater the precision of the variance estimate and thus, accuracy of projected sample sizes. Table 1 defines parameters of interest for the design and Figure 1 describes the general procedure. First, values are prespecified for target type I error rate (α_I), target power (P), the primary and secondary effect sizes of interest (δ_1 and δ_2), a curtailment bound (C_F) if futility stopping is included, and either the assumed true variance (σ^2) or the planning variance value (σ_0^2) depending on whether or not SSR based on accumulating data is considered (Figure 1). Next, for the treatment effect only design, sample sizes (n_1 and n_2), critical values for efficacy (C_{E1} and C_{E2}) can be calculated for both stages and the first stage subjects can be collected. Here the first stage sample size and critical value is based on powering for δ_1 and the second stage values are based on powering for δ_2 . With IP based SSR, only n_1 (powering for δ_1) can be calculated before the first stage is collected. Once the first stage data is collected, the SSR procedure will update both the effect size of interest and use the observed variance estimate, $\hat{\sigma}_1^2$, for determining n_2 , C_{E1} , and C_{E2} . Both procedures then follow the same pattern. The first stage test statistic is calculated and tested and the conditional power estimate, P_C is computed using first stage data and conditional on the true effect being the secondary (smaller) effect of interest. If the null hypothesis of no effect is not rejected ($Z_1 < C_{E1}$), and futility stopping bound is not achieved ($P_C < C_F$), then the second stage sample is collected and all data is used in a final test statistic for final testing.

3. Evaluation and Enumeration

3.1 Evaluating potential designs

There are a number of potential settings in which to consider and evaluate adaptive designs for CER trials. Potential design characteristics could include outcome variable type (e.g., continuous or binary), early stopping reason (futility, efficacy, both), SSR basis (effect size, variability, both), stopping bound type employed, and other underlying theoretical assumptions. Operating characteristics to evaluate for potential designs include the type I error rate, power, and expected sample size. Additionally, study characteristics can depend on the specific study parameters used for planning and that are assumed true. Extensive sensitivity analysis should be performed in order to assess the operating characteristics over a wide range of possibilities before an individual study is performed. We will consider a limited number of design variations in this manuscript to illustrate the potential of the design. In the Discussion (Section 4), we will describe additional scenarios and parameters of interest and that are being explored in forthcoming research results.

We believe that, unlike in typical trials with interim stopping bounds, which weight stopping bounds according to powering for a primary effect size of interest, the method described by Pocock (31) warrants consideration in a two-stage CER setting due to the equal significance level weighting it gives to the two potential stopping points. The method is more likely to allow stopping early for large effect size than the commonly employed OF (32) bounds. We

include and compare the OF (32) bounds with those described by Pocock (31). Alternative group sequential approaches like the triangular testing method and other two-stage methods (20–26,33,34) deserve review in future design evaluations. Fixed designs powered for the primary effect of interest and secondary effect of interest are also included for comparison purposes. Additional factors that we vary here include the true treatment difference, the true variance (as a ratio with the planning variance), and the use of conditional power based futility stopping rules.

3.2 HAM-D example

In order to illustrate the potential advantages of adaptive designs for use in CER trials, enumeration is presented for a subset of design possibilities as described in subsection 3.1. Recall the HAM-D example described in subsection 1.4. Here, two active treatments are being compared on the continuous outcome of 6 week decline in Hamilton Depression score. For all calculations, we assume a target type I error rate of $\alpha_I = 0.05$ and a target power of $P_I = 0.90$. As previously described, the effect size of interest is given as a range of $\delta = 1$ to 1.5 with the lower value being the smallest value that could affect first line clinical practice recommendations and the upper value being a treatment difference that the investigators would reasonably expect to see. The goal is an efficient study design that will declare significance as long as $\delta \geq 1$, with the understanding that there is good chance δ is as high as 1.5. The design is a two stage design with first stage designed to detect the larger primary effect ($\delta = 1.5$) and second stage designed to detect the smaller secondary effect ($\delta = 1$). The results were calculated allowing for futility or efficacy stopping at the first stage, or only allowing for efficacy stopping. For each scenario that allowed futility stopping, a conservative stopping bound of $C_F = 0.20$ was used. This means that those trials with less than a 20% chance of concluding efficacy, given the first stage data and assuming that the secondary (smaller) effect is true, were ended for futility.

Most of the results in this manuscript use exact theory developed for internal pilot with interim analysis designs (41). Use of the exact theory results in accurate calculations performed much faster than simulation studies would allow. The futility bound evaluations were conducted by simulation with 100,000 independent study replications per case. The calculations were performed with the SAS/IML software, Version 9.4 of the SAS System for Windows (Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA). For simplicity, all sample size and results calculations are based on large sample (Z -based) theory.

3.3 HAM-D Case 1: Known variance assumed

For the first enumeration, we assume that the variance is a known value ($\sigma^2 = 6$) and does not need to be re-estimated during the study. Also, we only consider early effectiveness stopping with interim SSR based on prespecified secondary effect size of interest. Calculations were made for true effect size: δ in $\{0, 1, 1.25, 1.5\}$. Results were calculated using α -spending functions (35) with the common O'Brien-Fleming (32) type stopping bounds and the bounds described by Pocock (31). Note that in this case without nuisance parameter based sample size re-estimation, this is a special case group sequential method.

The difference between this design and an ordinary GS design is that we would like high power to stop after the first stage if the expected effect size is reached. For comparison, Case 1 results include results for fixed sample designs having $n = 114$ and $n = 254$ subjects when powering for $\delta = 1.5$ or $\delta = 1$, respectively.

Power and sample size results for Case 1 can be seen in Table 2. Type I error rate inflation is low for all methods considered, but slightly higher in the Pocock method and the smaller of the fixed designs due to the use of Z-based critical values in small sample tests. The new methods do not introduce additional error above the fixed design in this setting. The addition of futility stopping without further adjustment causes the type I error rate to drop below the target under the null case of no effect. The addition of futility stopping also contributes to a 13% (OF) and 17% (Pocock) expected sample size savings under the null. The results for designs including futility stopping are virtually identical to those without them for all other cases considered and are not included separately in the Table 2. Both the Pocock and OF type stopping bounds achieve the desired power at $\delta = 1$, have high power for the intermediate effect size of $\delta = 1.25$, and virtually guarantee an effectiveness finding at $\delta = 1.5$. The fixed designs, as expected, perform at target levels for the effect sizes they are design for, but lose power for effects smaller than planned. For example, the design for $\delta = 1.5$ has power of 0.59 at $\delta = 1$. Large differences come, however, when comparing expected sample sizes. Both of the adaptive designs allow for sample size savings versus the fixed design at $\delta = 1$. However, the Pocock bounds allow for increased power to stop early while the OF bounds will rarely allow early stopping, translating to a 15–20% increase in average sample size for the OF designs.

3.4 HAM-D Case 2: Unknown variance

The second case is similar to the first in all aspects but one. Now we add the element of uncertainty concerning the true variance. In this situation, the planning variance is considered to be $\sigma_o^2 = 6$. However, we combine early effectiveness stopping with SSR based on the prespecified secondary effect size of interest *and* observed variance estimate from the first stage.

Table 3 gives power and expected sample size calculations for Case 2 using only Z-based calculations. The designs with smaller sample sizes exhibit some minor type I error rate inflation (up to 5.3%). The addition of futility, however, causes the type I error rate to be controlled in most settings without further adjustment. The addition of futility stopping also contributes to a significant expected sample size savings in all cases except when the true variance was larger than the pre-specified value. In this case, the additional second stage sample size determined by the SSR procedure increases the conditional power causing futility stopping to be difficult to achieve. Similar to Case 1, the results here for designs including futility stopping are virtually identical to those without them for all other cases and are not included separately in Table 3. As expected, the Pocock and OF methods have a much more stable power control than the fixed designs as shown in the $\delta = 1$, $\gamma = 2$ case where the variance was originally underestimated. Here the fixed designs have powers of 0.63 and 0.22 while the OF and Pocock methods have powers 0.90 and 0.89. The sample size benefits of the Pocock method become clear as the effect size increases due to the

chances of early stopping increasing with true effect size. Under the null, the OF method has a small sample size advantage over the Pocock method. In the Pocock and OF methods, if the originally specified variances value were too high, expected sample sizes decrease. Conversely, if the planning variance values were too low, the SSR techniques increase the second stage sample size to appropriate levels to create power stability. Adding futility stopping to the adaptive designs allows for a significant expected sample size savings under the null while not affecting power or sample size numbers over the other scenarios considered.

4. Discussion

The methods employed are adaptive in that they can lead to early stopping or resizing the study if it continues. They can protect against power loss from nuisance parameter underestimation while saving sample size if the opposite is the case. It is important to achieve alignment between design and goals during planning. If the goal is only to achieve power for effect sizes at (or above) a single point, and researchers are comfortable assuming nuisance parameters are known, then a fixed effect design is ideal. If, however, a range of effect sizes of interest is known beforehand, or there is considerable uncertainty concerning planning values for nuisance parameters (as is usually the case), then adaptive designs have much value.

There are ideas contained here that we believe advance clinical trial thinking. For one, the “preplanned grey zone” with a range of reasonable effects that should be accounted for in design consideration is an important idea in CER. Also, it is important to recall that small differences have more clinical meaning and more uncertainty of variance is likely in CER research. Accounting for both of these using traditional designs would result in very large and inefficient studies, which is exactly the opposite of the promise of CER. Adaptive designs have a lot of potential to allow CER trials to successfully and efficiently make important comparisons. In the preliminary results in the present study, the Pocock boundaries show some advantages compared to the OF boundaries. Before any recommendations can be made for their use, further comparisons using other stopping methods need to be undertaken.

Other additional work is needed to study the development and evaluation of new designs in CER trials in a multitude of potential settings. Our own ongoing work will include more extensive simulation results including continuous and binary outcomes as well as further work on efficiently incorporating futility stopping rules using conditional power arguments. We will also include a number of other stopping bound and testing strategies for two group, two stage comparisons including the use of triangular stopping bounds, minimax designs, and combination tests (20–26,33,34). Work to refine and automate methods of type I error rate control are also ongoing. Multiarm testing strategies are a natural extension of great interest for CER trials and such methods should be explored in the future (46–48). Additionally, Bayesian approaches to adaptive CER trials, which are not covered in this paper, are another interesting area of current research (2). One possibility in this context is for information from prior trials could be combined with accumulating information from a comparative trial to determine stopping thresholds or the dropping of arms during the course

of a study. Another is to incorporate Bayesian adaptive randomization in order to efficiently determine the best or worst treatment among a number of groups (27). The NHLBA funded Research in Adaptive methods for Pragmatic Trials (RE-ADAPT) project is a 'proof of concept' to stimulate further research into this area by assess whether Bayesian adaptive designs offer potential efficiencies in CER. They will do this by simulating a re-execution of the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) study using actual study data (28). A limitation to this work (and all work with sample size re-estimation and interim analysis) is the additional complexity when there is either a long delay between recruitment and outcome assessment, or a very short recruitment period. In either of these cases, a pause in recruitment, or continued recruitment and assessment during an interim analysis period must be critically evaluated for feasibility and usefulness of the adaptive procedure.

Conclusion

Traditional trial designs have a number of drawbacks in the modern CER trial setting. ADs in CER trials have the potential to allow trials to successfully and efficiently make important comparisons. We believe that through continued theoretical and enumerative research and discussion among the various interested parties, adaptive designs can unlock the potential of CER trials. It is imperative that these tools be known, developed, and available as CER rapidly moves into the front and center of our medical research attention.

Acknowledgments

We thank the two anonymous reviewers for their significant, detailed, and insightful contributions to the final manuscript.

References

1. Institute of Medicine. Initial national priorities for comparative effectiveness research. Washington, DC: Natl. Acad. Press; 2009. <http://www.nap.edu/catalog/12648.html>
2. Sox HC, Goodman SN. The methods of comparative effectiveness research. *Annu Rev Public Health*. 2012; 33:425–445. [PubMed: 22224891]
3. Steinbrook R. Health care and the American Recovery and Reinvestment Act. *N Engl J Med*. 2009; 360:1057–1060. [PubMed: 19224738]
4. Patient Protection and Affordable Care Act (2010). S. 6301, 111th Congress, 2nd Session. 2010
5. Patient-Centered Outcomes Research Institute. National priorities for research and research agenda [Internet]. Washington (DC): PCORI; 2012 May 2. Available from: <http://www.pcori.org/assets/PCORI-National-Priorities-and-Research-Agenda-2012-05-21-FINAL.pdf> Luce BR, Kramer JM, Goodman SN, et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Ann Intern Med*. 2009; 151:206–209. [PubMed: 19567619]
6. Tunis SR, Benner J, McClellan M. Comparative effectiveness research: policy, context, methods development, and research infrastructure. *Stat Med*. 2010; 29:1963–1976. [PubMed: 20564311]
7. Sullivan P, Goldman D. The promise of comparative effectiveness research. *JAMA*. 2011; 305:400–401. [PubMed: 21266687]
8. Chalkidou K, Tunis S, Whitcher D, et al. The role for pragmatic randomized controlled trials (pRCTs) in comparative effectiveness research. *Clin Trials*. 2012; 9:436–446. [PubMed: 22752634]
9. Zwarenstein M, Treweek S, Gagnier J, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ*. 2008; 337:a2390. [PubMed: 19001484]

10. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic–explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol*. 2009; 62:464–475. [PubMed: 19348971]
11. Davis BR, Cutler JA, Gordon DJ, et al. Rationale and design for the antihypertensive and lipid lowering treatment to prevent heart attack trial (ALLHAT). *Am J Hypertens*. 1996; 9:342–360. [PubMed: 8722437]
12. Davis BR, Cutler JA, Gordon DJ. Major outcomes in high risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid Lowering treatment to prevent Heart Attack Trial (ALLHAT). *JAMA*. 2002; 288:2981–2997. [PubMed: 12479763]
13. Stafford RS, Monti V, Furberg CD, Ma J. Long-term and short-term changes in antihypertensive prescribing by office-based physicians in the United States. *Hypertension*. 2006; 48:213–218. [PubMed: 16785334]
14. Morton SC, Ellenberg JH. Infusion of statistical science in comparative effectiveness research. *Clin Trials*. 2012; 9:6–12. [PubMed: 22334463]
15. Food and Drug Administration. Guidance for industry: adaptive design clinical trials for drugs and biologics draft guidance. 2010. Accessed at <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm>
16. Gallo P, Chuang-Stein C, Dragalin V, et al. Adaptive designs in clinical drug development: an executive summary of the PhRMA working group. *J Biopharm Stat*. 2006; 16:275–283. [PubMed: 16724485]
17. Kairalla JA, Coffey CS, Thomann MA, Muller KE. Adaptive trial designs: a review of barriers and opportunities. *Trials*. 2012; 13:145. [PubMed: 22917111]
18. Lai TL, Lavori PW, Shih MC. Adaptive trial designs. *Annual review of pharmacology and toxicology*. 2012; 52:101–110.
19. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Stat Med*. 1999; 18:1833–1848. [PubMed: 10407255]
20. Lehman W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999; 55:1286–1290. [PubMed: 11315085]
21. Bretz F, Koenig F, Brannath W, et al. Adaptive designs for confirmatory clinical trials. *Stat Med*. 2009; 28:1181–1217. [PubMed: 19206095]
22. Whitehead J, Valdes-Marquez E, Lissmats A. A simple two-stage design for quantitative responses with application to a study in diabetic neuropathic pain. *Pharm Stat*. 2009; 8:125–135. [PubMed: 18642403]
23. Wason JM, Mander AP. Minimizing the maximum expected sample size in two-stage phase II clinical trials with continuous outcomes. *J Biopharm Stat*. 2012; 22:836–852. [PubMed: 22651118]
24. Menon S, Chang M. Optimization of adaptive designs: efficiency evaluation. *J Biopharm Stat*. 2012; 22:641–661. [PubMed: 22651106]
25. Liu Q, Li G, Anderson KM, Lim P. On efficient two-stage adaptive designs for clinical trials with sample size adjustment. *J Biopharm Stat*. 2012; 22:617–640. [PubMed: 22651105]
26. Connor JT, Elm JJ, Broglio KR. Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus. *J Clin Epidemiol*. 2013; 66:S130–S137. [PubMed: 23849147]
27. Connor JT, Luce BR, Broglio KR, et al. Do Bayesian adaptive trials offer advantages for comparative effectiveness research? Protocol for the RE-ADAPT study. *Clin Trials*. 2013; 10:807–827. [PubMed: 23983160]
28. Tsiatis AA, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*. 2003; 90:367–378.
29. Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. *Stat Med*. 2006; 25:917–932. [PubMed: 16220524]
30. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977; 64:191–199.

31. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35:549–556. [PubMed: 497341]
32. Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. *Biometrics*. 1983; 39:227–236. [PubMed: 6871351]
33. Whitehead J, Todd S. The double triangular test in practice. *Pharm Stat*. 2004; 3:39–49.
34. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983; 70:659–663.
35. Jennison, C.; Turnbull, BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: Chapman & Hall/CRC; 2000.
36. Cui L, Hung HMJ, Wang S. Modifications of sample size in group sequential clinical trials. *Biometrics*. 1999; 55:853–857. [PubMed: 11315017]
37. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Stat Med*. 1990; 9:65–72. [PubMed: 2345839]
38. Mehta CR, Tsiatis AA. Flexible sample size considerations using information-based interim monitoring. *Drug Inf J*. 2001; 35:1095–1112.
39. Tsiatis AA. Information-based monitoring of clinical trials. *Stat Med*. 2006; 25:3236–3244. [PubMed: 16927248]
40. Kairalla JA, Muller KE, Coffey CS. Combining an internal pilot with an interim analysis for single degree of freedom tests. *Commun Stat Theory Methods*. 2010; 39:3717–3738. [PubMed: 21037942]
41. Kairalla JA, Coffey CS, Muller KE. Achieving the benefits of both an internal pilot and interim analysis in large and small samples. 2010 JSM Proceedings, ENAR Section. 2010:5239–5252.
42. Miller F. Variance estimation in clinical studies with interim sample size reestimation. *Biometrics*. 2005; 61:355–361. [PubMed: 16011681]
43. Mehta CR, Patel NR. Adaptive, group sequential, and decision theoretic approaches to sample size determination. *Stat Med*. 2006; 25:3250–3269. [PubMed: 16927402]
44. Hamilton M. Rating depressive patients. *J Clin Psychiatry*. 1980; 41:21–24. [PubMed: 7440521]
45. Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. *Biometrika*. 1988; 75:303–310.
46. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Stat Med*. 2003; 22:689–703. [PubMed: 12587100]
47. Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika*. 2012; 99:494–501.
48. Wason JMS, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials. *Stat Methods Med Res*. 2012 Published online December 12, 2012.

Figure 1.
The General Procedure
Dashed line and C_F are for futility designs only

Table 1

Symbols of Interest for Adaptive CER Trials

Symbol	Definition
α_t	Target type I error rate
P_t	Target power
n_1	First stage sample size
n_2	Second stage sample size
C_{E1}	First stage effectiveness bound
C_{E2}	Second stage effectiveness bound
C_F	First stage futility bound
P_C	Conditional power
δ	True mean difference
δ_1	Primary effect of interest
δ_2	Secondary effect of interest
σ^2	True residual variance
σ_0^2	Planning residual variance
$\hat{\sigma}_1^2$	First stage observed variance estimate

Table 2

Power and expected sample size for HAM-D Case 1

True δ	Fixed ($\delta = 1$)		Fixed ($\delta = 1.5$)		O'Brien-Fleming		Pocock	
	Power $\times 100$	N	Power $\times 100$	N	Power $\times 100$	E(N) [#]	Power $\times 100$	E(N) [#]
0	5.1	254	5.2	114	5.1	255	5.2	273
0 (w/futility [*])	-	-	-	-	3.1	221	4.3	227
1	90.2	254	58.8	114	90.3	224	89.8	195
1.25	98.2	254	77.7	114	98.3	197	98.2	164
1.5	99.8	254	90.4	114	99.8	167	99.8	141

* Conditional power based futility stopping at first stage with stopping bound at $C_F=0.2$ under secondary effect size assumption. For all other true effect sizes δ the power and sample size differences with adding futility stopping were negligible.

[#] Maximum sample size for the O'Brien-Fleming and Pocock designs is 256 and 278, respectively

Table 3

Power and expected sample size for HAM-D Case 2

True δ	$\gamma = \sigma^2/\sigma^2_0$	Fixed ($\delta = 1$)			Fixed ($\delta = 1.5$)			O'Brien-Fleming		Pocock	
		Pow $\times$ 100	N	N	Pow $\times$ 100	N	N	Pow $\times$ 100	E(N) [#]	Pow $\times$ 100	E(N) [#]
0	0.5	5.1	254	254	5.2	114	114	5.3	129	5.3	141
	1	5.1	254	254	5.2	114	114	5.3	254	5.3	274
	2	5.1	254	254	5.2	114	114	5.1	509	5.2	544
0 (w/futility [*])	-	-	-	-	-	-	-	4.4	116	4.9	126
	-	-	-	-	-	-	-	3.4	222	4.4	231
	-	-	-	-	-	-	-	5.0	506	5.2	543
1	0.5	99.6	254	254	63.8	114	114	90.1	118	92.4	127
	1	90.2	254	254	58.8	114	114	89.9	225	89.8	198
	2	63.3	254	254	21.7	114	114	89.8	505	88.7	440
1.25	0.5	>99.9	254	254	97.0	114	114	98.3	115	98.9	125
	1	98.2	254	254	77.7	114	114	98.0	201	98.1	165
	2	82.0	254	254	48.9	114	114	98.0	501	97.8	378
1.5	0.5	>99.9	254	254	99.6	114	114	99.8	114	99.9	124
	1	99.8	254	254	90.4	114	114	99.8	172	99.8	142
	2	93.1	254	254	63.8	114	114	99.7	491	99.7	310

* Conditional power based futility stopping at first stage with stopping bound at $C_T=0.2$ under secondary effect size assumption. For all other true effect sizes δ and true variance ratios γ , the power and sample size differences with adding futility stopping were negligible.

Achieved sample size for the SSR designs depends on the continuous observed variance estimate. For $\gamma = 0.5$, 1, and 2, the near-maximum sample sizes based on 99th percentile of the variance estimate is 170, 340, and 680 for the O'Brien-Fleming designs, and 184, 368, and 734 for the Pocock designs, respectively.