Adaptive Designs for Comparative Effectiveness Trials

John A. Kairalla[1], Christopher S. Coffey[2], Mitchell A. Thomann[2]
Keith E. Muller[3]
[1]Department of Biostatistics, University of Florida, PO Box 117450, Gainesville, FL 32611-7450, USA
[2]Department of Biostatistics, University of Iowa, 2400 University Capitol Centre, Iowa City, IA 52240-4034, USA
[3]Department of Health Outcomes and Policy, University of Florida, PO Box 100177, Gainesville, FL 32610-0177, USA

Abstract

There is a clearly increasing level of national interest in comparative effectiveness research (CER). Due to the newness of the field and also confusion about what exactly CER is, reliable evidence based methods on the comparative effectiveness (CE) of treatments for medical and health policy decision makers are currently inadequate. RCTs must have a prominent place in CER due to their reliable information and well respected standard. However, in randomized CE trials, there may be limited information to guide initial design choices including the patient population, the primary outcome, or the target effect size. In the general RCT setting, adaptive designs have been proposed to address these concerns. There are potential advantages to expanding adaptive designs to within the CE context. Although there are many similarities between the two, CE trials have some fundamental differences from standard clinical trials. For one, the heterogeneity in the population studied in CE creates higher variability in outcomes. CE studies could be underpowered if they use planning values obtained from tightly controlled clinical trials. Additionally, the concept of a 'minimum clinically meaningful difference' is hard to define in the CE context. Even assuming equal cost and safety, a range of meaningful effect sizes could be defined with upper limit the largest effect with reasonable chance of being observed and lower limit the minimal effect deemed sizable enough to change practice in the study context. We first review the current state of clinical CER. Then, we identify areas of CER that seem particularly strong candidates for the development of novel adaptive design methodology and application. We describe the evaluative process to determine the usefulness of these designs in CER in a number of useful two group comparison situations. Illustrative analytic results are used to explore properties of various adaptive sample size re-estimation designs tailored for use in CE trials. We summarize results, make recommendations, and identify areas needing future research.

Key Words: adaptive designs, comparative effectiveness; sample size re-estimation; power analysis

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 in the United States typically depend on narrowly defined *efficacy* endpoints. That is, they 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 is 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 in the United States 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, often, many such potential treatments 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 in traditional designs.

1.2 Comparative Effectiveness Research

The field of comparative effectiveness research (CER) has grown as a response to the costs and drawbacks of traditionally designed research designs. 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 [1]. 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 recent national healthcare reform legislated the creation of a national Patient-Centered Outcomes Research Institute to guide expansion of CER [2,3].

An Institute of Medicine Committee on Initial Priorities for CER [4] 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 on the "generation...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 are currently inadequate [5].

Tunis et al. [6] provided an excellent summary of the current policy context, need for future methods development in CER research, and summary of existing research infrastructure. They also clarify 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 of interest, the focus of this manuscript lies with the use of RCTs for conducting CER. CE research can create a general population that can fairly compare multiple treatments that were originally researched in very specific and differing populations and only compared versus placebo. 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 [5,7]. Additionally, unique issues must be addressed in order achieve the promised benefits of CER trials.

1.3 Motivating Example

As a motivating example, consider the planning for a two arm randomized controlled trial comparing active depression treatments with similar safety, and availability (adapted from [8]). The primary endpoint for the hypothesized study is 6 week decline in the Hamilton Depression Index (HAM-D; [9]). As both treatments being considered are proven active treatments, there is general uncertainty about the expected mean group-wise difference ($\delta = \mu_1 - \mu_2$). On one hand, based on their personal beliefs, the investigators believe that $\delta = 4$ is a reasonable expectation, but also recognize that any true value above $\delta = 2$ would be clearly clinically 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 appropriately determine the sample size for the study. For example, with respect to the mean group-wise difference (δ), should the reasonable expectation ($\delta = 4$) be used to plan a smaller study with good chance of success for higher effect sizes? Or should a much larger study be planned, assuming $\delta = 2$, in order to ensure that success is achieved for smaller values? The latter case would call for four times the upfront sample size commitment, using much more resources and time. In our experience, the larger but reasonable effect size expectation ($\delta = 4$) would be the most common choice made during planning.

Consider possible research outcomes under this scenario. The best case would be obtaining a result with one treatment clearly better than the other ($\hat{\delta} > 4$). Here the design was appropriate and the study gives a largely definitive answer to the research question. Another potential outcome would be finding very little difference between the two treatments ($\hat{\delta} < 2$). Here, the treatments may be considered to more or less have 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. A more disappointing result in this setting would be one with an observed treatment difference falling between two and four (e.g., $\hat{\delta} = 3$). Here, there is no statistically significant difference, but there could be a clinically meaningful difference present. We refer to this area as the ‘statistical gray zone’. As there is no fallback position in a CE 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.

Another potential issue when using traditional study design is that there is often uncertainty during planning about variance, σ^2. If this value is underestimated, the result is an underpowered study. Alternatively, an overestimate contributes to study inefficiency through increased sample size.

2. Adaptive Comparative Effectiveness Trials

2.1 Some issues with Comparative Effectiveness Trials

Valid and meaningful trials for CER are difficult to design for a number of reasons. For one, the concept of a 'minimum clinically meaningful difference' has diminished meaning in CE 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. Additionally, smaller effect sizes are expected in CE trials comparing proven treatments. Designs using traditional methods would require large sample sizes or only be powered to detect large effect sizes. Thus, clinically small, but population important, differences may be missed. Finally, heterogeneity in the 'real-world' populations studied in CER creates potentially higher variability in outcomes. Unreliable prior information could only be available from highly controlled studies from homogeneous populations.

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 Adaptive Designs

Adaptive designs (AD) methods continue to attract substantial interest in regulatory health science, as evidenced by the recently released U.S. Food and Drug Administration (FDA) draft guidance document [10]. ADs give one way to address the uncertain choices that must be made during planning for a CE 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 [11] stated that adaptive designs "…modify aspects of the study as it continues, without undermining the validity and integrity of the trial." Additionally, they state 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 more efficient treatment comparisons by reducing trial size and time, and by increasing the chance of a trial correctly answering the question of interest. More information on adaptive designs in general can be found in a recent review by Kairalla et al. [12]. Of particular interest for use in CE 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 different sources of 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 mis-specification. 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 [13] and by O'Brien and Fleming [14]. The method by Pocock takes the approach of finding a single adjusted nominal significance level that can be used at each testing time. Alternatively, the 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. We believe, however, that the method described by Pocock [13] holds promise in the CER setting and should be considered during study planning and design evaluation. Additionally, α-spending functions [15] are important tools that allow flexible timing of analyses regardless of the type I error preservation method employed. For futility stopping consideration, a common method in group sequential methods that seems promising in CER is the conditional power approach to stochastic curtailment [16]. GS methods in general are well known and have been extensively described [16].

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 based on observed effect size has generated considerable discussion and controversy [17-19] 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 internal pilots (IP; [20]) are two stage designs with no interim testing, but with interim sample size re-estimation based on first stage nuisance parameter estimates. The designs protect study power at the preplanned clinical effect of interest. Also, an IP design implemented in a setting where non-objective parties do not have access to accumulating raw data will give no information concerning effect trends of interest [12].

Considerable focus has been put on combining GS and IP based SSR designs in order to simultaneously achieve their benefits. Asymptotically correct methods for use of GS and IP methods in large clinical trials have been proposed [21,22]. These methods, however, can have type I error rate inflation in small samples. Exact distributional theory for internal pilots with interim analysis (IPIA) has computational time advantages over simulation methods and power and expected sample size benefits over fixed sample methods [23]. Kairalla, Coffey, and Muller [24] identified three sources of type I error rate inflation in IPIA designs and showed how they could be effectively controlled.

Current and ongoing research seeks to incorporate these methods into potential CE trial designs in order to efficiently allow for sample size and study duration flexibility. A selection of results showing their promise is included in section 3.

2.2 A Potential Adaptive CE Trial

A proposed new aspect of an adaptive CE trial would be the introduction of a primary and secondary effect size on 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 change practice in a study context. Thus, an example AD for two group CE trials could have two stages with the first powered to detect the larger reasonable effect size (such as 4 points in HAM-D reduction). At the conclusion of the first stage, one of three decisions could be reached: 1) Declare efficacy (one treatment clearly better), 2) Declare futility (study unlikely to show difference between treatments), or 3) If results suggest a smaller effect might exist, then proceed with a second stage powered to detect the smaller effect. 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 nuisance parameters are also uncertain, additional SSR calculations for the second stage could also incorporate observed nuisance information based on the first stage data. An example for continuous outcomes would be the use of the observed variance estimate at the interim stage in determining the second stage sample size.

2.3 Evaluating Potential Designs

There are a number of potential settings in which to consider adaptive designs for CE trials settings. Additionally, for each study setting, there are many potential design variations to consider when evaluating the operating characteristics and robustness of a design. For example, study settings and variations could include combinations of the following:

- Outcome variable type: continuous, binary, or time-to-event
- Early stopping reason: futility, effectiveness, or both
- Sample size re-estimation reason: effect size, nuisance parameter, or both
- Stopping bound type employed: Pocock, OF, conditional power
- Theory used for critical values and power calculations: large sample (*Z*) or small sample (*t*) theory

Operating characteristics to evaluate include the type I error rate, power, and expected sample size. These values depend not only on the design considerations above, but also 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. Table 1 gives an idea of the parameters of interest that should be examined for a given study type comparing two treatments with either a continuous or binary outcome, SSR based on treatment effect or treatment effect and nuisance parameter, and for early stopping abilities for efficacy or for efficacy and futility.

Table 1: Parameters of Interest for Sensitivity Analysis

		Sample size re-estimation based on:	*Treatment*		*Trt and Nuisance*	
		Possible first stage stopping conclusions:	*Eff*	*Eff or Fut*	*Eff*	*Eff or Fut*
	Symbol	*Definition*				
Both	α_t	Target type I error rate	X	X	X	X
Outcomes	P_t	Target power	X	X	X	X
	n_1	Planned first stage sample size	X	X	X	X
	n_2	Planned second stage sample size	X	X		
	$n_{+,\mathrm{max}}$	Maximum allowed size of final sample			X	X
	C_{E1}	First stage effectiveness bound (stop if $\|Z\| \geq C_{E1}$)	X	X	X	X
	C_{F1}	First stage futility bound (stop if $\|Z\| \leq C_{F1}$)		X		X
	C_{E2}	Second stage effectiveness bound	X	X	X	X
Continuous	δ	True mean difference	X	X	X	X
Outcomes	δ_1	Primary effect of interest	X	X	X	X
	δ_2	Secondary effect of interest	X	X	X	X
	σ^2	True variance			X	X
	σ_0^2	Variance value for planning			X	X
Binary	ϵ	True proportion difference	X	X	X	X
Outcomes	ϵ_1	Primary effect of interest	X	X	X	X
	ϵ_2	Secondary effect of interest	X	X	X	X
	$\overline{\pi}$	True pooled event rate			X	X
	$\overline{\pi}_0$	Pooled event rate for planning			X	X

3. Enumeration

3.1 HAM-D Example

In order to exemplify the potential advantages of adaptive designs for use in CE trials, enumeration will be presented for a limited subset of the possibilities listed in section 2. Recall the HAM-D example described in subsection 1.3. 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 $\alpha_t = 0.05$ and a target power of $P_t = 0.90$. As previously described, the effect size of interest is given as a range of δ = 2 to 4 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 $\delta \geq 2$, with the understanding that there is good chance δ is as high as 4. The design is a two stage design with first stage designed to detect larger effect (4) and second stage designed to detect smaller effect (2).

While more complex situations would typically use simulations for sensitivity analyses, the results in this manuscript all use exact theory developed for internal pilot with interim analysis designs [23]. Use of the exact theory results in accurate calculations performed much faster than simulation studies would allow. The calculations were performed with the SAS/IML software, Version 9.2 of the SAS System (Copyright © 2002-2008, 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).

3.2 HAM-D Case 1: Known Variance Assumed

For the first enumeration, we will assume that the variance is a known value ($\sigma^2 = 4$) and does not need to be re-estimated during the study. Also, we will only consider early effectiveness stopping with interim SSR based on secondary effect size of interest. Calculations will be made for true effect size: δ in {0, 2, 4}. Results will be calculated using α-spending functions [15] with the common O'Brien-Fleming [14] type stopping bounds and the bounds described by Pocock [13]. The example will result in a fairly small sample setting, with a fixed sample design having n=44 subjects when powering for $\delta = 2$. We will compare the large sample theory (*Z*-based) with the more exact *t*-based methods. Note that in this case without nuisance parameter based sample size re-estimation, we have 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 our expected effect size is reached. Calculation for this example is based on exact theory.

Table 2: Power and expected sample size for HAM-D Case 1
(using exact theory calculations)

		Z-based		*t-based*	
Bound type	*True δ*	*Power x 100*	*E(N)*	*Power x 100*	*E(N)*
Fixed (δ = 2)	0	5.7	44	5.0	46
O'Brn-Flem	0	5.9	43.9	5.0	46.0
Pocock	0	7.3	46.4	5.1	49.1
O'Brn-Flem	2	91.0	41.9	91.2	45.5
Pocock	2	90.3	34.5	90.0	38.0
O'Brn-Flem	4	>99.9	29.8	>99.9	38.6
Pocock	4	>99.9	16.5	>99.9	18.6

Power and sample size results for Case 1 can be seen in Table 2. Note that all of the bound types have unacceptable type I error rate inflation (power at $\delta = 0$) when using the large sample Z-based critical value and sample size calculations. However, the t-based values better account for the uncertainty in the variance estimate and appropriately control the type I error rate. Both the Pocock and OF type stopping bounds achieve the desired power at $\delta = 2$ and virtually guarantee an effectiveness finding at $\delta = 4$. The difference comes, however, when comparing expected sample sizes. When $\delta = 2$, the Pocock bounds allow for increased power to stop early while the OF bounds will rarely allow early stopping, translating to an average 20% additional sample size. The effect is much more dramatic at the top of the effect range of interest. Here, the Pocock bounds translate to less than half the sample size compared to the OF bounds. Note that no futility stopping is included here, resulting in no sample size savings versus the fixed sample design under no effect ($\delta = 0$).

3.3 HAM-D Case 2: Unknown Variance

The second case will be 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 of change is considered to be $\sigma^2_o = 4$. However, we will combine early effectiveness stopping with SSR based on secondary effect size of interest *and* observed variance estimate from the first stage. Since in the previous example we showed the benefits of t-based methods and Pocock bounds for this situation, we will constrain our results to this setting. To help control type I error rate inherent to sample size re-estimation designs, we will consider an approximate *bounding method* (Coffey and Muller, 2001) modified from use in internal pilot designs. With α_t the target type I error rate, the actual type I error rate depends on the ratio of the true variance to the planning value, $\gamma = \sigma^2/\sigma^2_0$. The bounding method finds nominal $\alpha_b \leq \alpha_t$, such that the test has type I error rate no more than α_t over all possible γ. Work is needed on finalizing a numeric algorithm to automatically calculate α_b in this context. A few trial and error calculations gave $\alpha_b = 0.0465$ as a reasonable value to use here for illustrative purposes.

Table 3: Power x 100 for HAM-D Case 2: t-based with Pocock bounds (using exact theory calculations)

True δ	$\gamma = \sigma^2/\sigma^2_o$	*Fixed*	*Pocock*	*Pocock-Bound*
0	0.5	5.0	5.4	5.0
0	1	5.0	5.4	5.0
0	2	5.0	5.2	4.8
2	0.5	>99	93.3	93.2
2	1	91.2	89.2	89.1
2	2	65.0	86.5	86.4
4	0.5	>99	>99	>99
4	1	>99	>99	>99
4	2	>99	>99	>99

Table 3 gives power calculations for Case 2 using only t-based calculations and Pocock like stopping bounds, both with and without the bounding method. The Pocock without the bounding method exhibits some type I error rate inflation. However, the Pocock-Bounding method has virtually the same power results, but with type I error rate control. The Pocock methods have a much more stable power control than the fixed design as shown in the $\delta = 2, \gamma=2$ case where the variance was originally underestimated. Here the fixed design has power of 0.65 while the Pocock methods are over 0.86.

Table 4: Expected sample size for HAM-D Case 2: *t*-based with Pocock bounds (using exact theory calculations)

True δ	$\gamma = \sigma^2 / \sigma^2_o$	*Fixed*	*Pocock*	*Pocock-Bound*
0	0.5	46	28	28
0	1	46	50	51
0	2	46	95	97
2	0.5	46	21	22
2	1	46	41	42
2	2	46	86	88
4	0.5	46	16	16
4	1	46	20	20
4	2	46	51	53

Table 4 displays the expected sample size information for Case 2. As expected, 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. Also, if the originally specified variance value was too high, expected sample size decreases. Conversely, if the planning variance was too low, the SSR technique increases the second stage sample size to appropriate levels to create power stability.

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 a single effect size point, and 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, then adaptive designs have much value.

There are a few ideas contained here that are new to modern clinical trial thinking. For one, the 'preplanned grey zone' with a range of reasonable effects that should be accounted for 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 in this area in order to allow CE trials to successfully and efficiently make important comparisons. A new finding is that this seems to be a rare area where Pocock bounds seem appropriate and work quite well. Further work and discussion should seriously consider their implementation in such designs. Work to refine and automate methods of type I error rate control are ongoing. Other additional work is needed to study the development and evaluation of new designs in CE trials in a multitude of potential settings. Also, Bayesian approaches to adaptive CE trials have been discussed somewhat in the literature [1], but are not addressed here. If successfully implemented, they could help incorporate prior information into a study in order to boost efficiency.

We believe that through continued theoretical and enumerative research and discussion among the various interested parties, adaptive designs can unlock the potential of CE 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.

Acknowledgements

We gratefully acknowledge the advice and assistance of our colleagues Dr. Ronald Shorr at the University of Florida and the Malcom Randall VA Medical Center, and James Toner and Elizabeth Chrischilles from the University of Iowa.

All authors were supported in part by a supplement to the NIH/NCRR Clinical and Translational Science Award to the University of Florida, NCRR UL1RR029890-03S1. Additional support for JAK included NINR 1 R01 AG039495-01. Additional support for CSC included NINDS U01-NS077352, NINDS U01-NS077108, NINDS U01-NS038529, and NHLBI R01-HL091843-04. Additional support for KEM included NIDDK R01-DK072398, NIDCR U54-DE019261, NIDCR R01-DE020832-01A1, NHLBI R01-HL091005, NIAAA R01-AA016549, and NIDA R01-DA031017.

References

1. Sox HC, Goodman SN (2012). The methods of comparative effectiveness research. *Annual Review of Public Health*, **33**:425-445.
2. Steinbrook R (2009). Health care and the American Recovery and Reinvestment Act. *New England Journal of Medicine*, **360**(11):1057-1060.
3. Patient Protection and Affordable Care Act (2010). S. 6301, 111th Congress, 2nd Session, 2010.
4. Institute of Medicine (2009). Initial national priorities for comparative effectiveness research. Washington, DC: Natl. Acad. Press. http://www.nap.edu/catalog/12648.html.
5. Luce BR, Kramer JM, Goodman SN, Connor JT, Tunis S, Whicher D, Schwartz JS (2009). Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Annals of Internal Medicine*, **151**:206-209.
6. Tunis SR, Benner J, McClellan M (2010). Comparative effectiveness research: policy, context, methods development, and research infrastructure. *Statistics in Medicine*, **29**:1963-1976.
7. Sullivan P, Goldman D (2011). The promise of comparative effectiveness research. *Journal of the American Medical Association*, **305**(4):400-401.
8. Mehta CR, Patel NR (2006). Adaptive, group sequential, and decision theoretic approaches to sample size determination. *Statistics in Medicine*, **25**:3250-3269.
9. Hamilton M (1980). Rating depressive patients. *Journal of Clinical Psychiatry***,** **41**:21-24.
10. Food and Drug Administration (2010). Guidance for industry: adaptive design clinical trials for drugs and biologics draft guidance. Accessed at http:/www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm
11. Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J (2006). Adaptive designs in clinical drug development: an executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics*, **16**:275–283.

12. Kairalla JA, Coffey CS, Thomann MA, Muller KE (2012). Adaptive trial designs: a review of barriers and opportunities. *Trials*, **13**:145.
13. Pocock SJ (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**:191-199.
14. O'Brien PC, Fleming TR (1979). A multiple testing procedure for clinical trials. *Biometrics*, **35**:549–556.
15. Lan KKG, DeMets DL (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, **70**:659-663.
16. Jennison C, Turnbull BW (2000). Group Sequential Methods with Applications to Clinical Trials. Boca Raton: Chapman & Hall/CRC.
17. Cui L, Hung HMJ, Wang S (1999). Modifications of sample size in group sequential clinical trials. *Biometrics*, **55**:853-857.
18. Tsiatis AA, Mehta C (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, **90**:367-378.
19. Jennison C, Turnbull BW (2006). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine*, **25**:917-932.
20. Wittes J, Brittain E (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, **9**:65-72.
21. Mehta CR, Tsiatis AA (2001). Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal*, **35**:1095-1112.
22. Tsiatis AA (2006). Information-based monitoring of clinical trials. *Statistics in Medicine*, **25**:3236-3244.
23. Kairalla JA, Muller KE, Coffey CS (2010). Combining an internal pilot with an interim analysis for single degree of freedom tests. *Communications in Statistics-Theory and Methods*, **39**(20):3717-3738.
24. Kairalla JA, Coffey CS, Muller KE (2010). Achieving the benefits of both an internal pilot and interim analysis in large and small samples. *2010 JSM Proceedings, ENAR Section*, 5239-5252.