Statistica Sinica **23** (2013), 1775-1791
doi:http://dx.doi.org/10.5705/ss.2012.231s

SEQUENTIAL METHODS FOR COMPARATIVE EFFECTIVENESS EXPERIMENTS: POINT OF CARE CLINICAL TRIALS

Mei-Chiung Shih[1,2] and Philip W. Lavori[1]

[1]*Stanford University and*
[2]*VA Palo Alto Cooperative Studies Program Coordinating Center*

Abstract: The goal of comparative effectiveness research (CER) is to support evidence-based choices of treatments. Currently the majority of randomized trials for CER are designed to demonstrate superiority, which often require large sample size because the effect sizes between treatments in current use are typically small to moderate and there are usually more than two treatments to be compared. We propose an alternative group sequential design for such setting. Instead of testing superiority, we aim to select high quality treatments that are within a small distance from the best treatment. The basic idea is to eliminate non-promising treatments at interim analyses that cannot be much better than the currently observed best treatment, based on generalized likelihood ratio tests. This approach can also be used for guideline implementation and for phase II selection trials.

Key words and phrases: Comparative effectiveness research, generalized likelihood ratio, guideline implementation, phase II-III trials, point of care clinical trials, sequential design.

1. Introduction

Comparative Effectiveness Research (CER) involves the comparison of medical, surgical, and other treatments on their ability to benefit patients. The word ‘effectiveness’ conveys the sense that the treatments are to be compared on their effects as they are actually used by clinicians and received by patients, in patient populations selected by diagnosis and other relevant considerations, but otherwise not restricted by race, sex, economic condition, insurance coverage, likelihood of successful adherence to treatment, and the like. Effectiveness contrasts with ‘efficacy’, a term usually denoting the presence or absence of a desired effect in somewhat ideal circumstances of adherence to treatment and often in more highly selected populations (those without serious co-morbid conditions, in a restricted range of severity, or otherwise selected for homogeneity). CER is often approached by observational methods, including analysis of claims data or registry data. For example, Stukel et al. (2007) describe several attempts to use statistical analysis of large Medicare claims databases to compare survival rates after medical and surgical treatments for acute cardiovascular disease. The key problem with observational approaches involves 'confounding by indication', the tendency for freely choosing clinicians and patients to choose treatments with their anticipated effects in mind. Such 'non-ignorable' treatment assignment (Rubin (1974, 1978)) creates the possibility of bias in the estimation of effectiveness, which is dealt with by statistical adjustment and modeling techniques, or instrumental variables methods, or some combination.

Randomization offers another way to remove confounding by indication, leading to experimental CER (E-CER). The cost and complexity of randomized trials compares unfavorably with the apparent ease of analysis of observational data, so E-CER represents a small fraction of the totality of CER. A criticism of the traditional randomized clinical trial approach to CER is that it takes too long, and by the time its results are available, they are out of date. Recent advances in the infrastructure of clinical experiments in natural clinical settings, such as Point of Care Clinical Trials (POC-CT) (Fiore et al. (2011)), offer the possibility of using E-CER more easily and at lower cost, obtaining more reliable information to guide clinical decision-making. In this article, we assume wider use of such advances in informatics and decision support for trials, which we believe will enable a new generation of rapid E-CER with fewer cost constraints. We propose a novel framework for design of such trials that takes into account some design goals and constraints of E-CER that we describe below.

The usual 'superiority' design used in traditional E-CER is somewhat confounded by the fact that CER compares (by definition) treatments in current use, with no strong basis for preferring one over the others. In contrast to the strong hopes that attach to new treatments during their development, large effect sizes may be implausible in E-CER. Such prior limits on the plausible size of anticipated treatment effects drive sample sizes up in the usual inverse-quadratic way. So-called non-inferiority designs offer an alternative, but have their own limitations, some having to do with the asymmetry of the typical design because non-inferiority is usually stated relative to a specified standard treatment of known effectiveness. A more practical and useful goal in E-CER than hypothesis testing of superiority or non-inferiority is to select a treatment that is non-inferior to the unknown best treatment.

Recent discussions of the need for healthcare systems to 'learn' more effectively reveal a new arena for clinical trials designs motivated by treatment selection. In a large accountable care organization (ACO), there may be little incentive to pay the price for testing superiority, but a great deal of interest in ensuring that substantially inferior treatments are systematically weeded out. The control of type I error in superiority testing is of primary value to regulators such as the U.S. Food and Drug Administration (FDA), while ACO administrators may not need to demonstrate superiority of the best among the treatments in use.

Decision-theoretic approaches to design of trials take into account the finite number of future patients for whom the trial results will be definitive. There are two contributions to the loss function: the patients treated during the trial with the inferior treatment, and the patients in the future who are treated with the inferior treatment because the trial comes to an incorrect decision. The tradeoff between getting the right answer in the trial and capitalizing on that answer in the remaining patient population is known as the 'patient horizon problem'. Along with the usual desire to limit patient exposure in trials to relatively ineffective treatments, the decision-theoretic point-of-view motivates designs that adapt the randomization to sequential estimates of treatment differences. The Bayesian designs of Berry (2010, 2012), Berry et al. (2011), Huang et al. (2009), various 'play the winner' methods, and others provide designs that are attractive from an ethical point of view. These designs may ignore the operating characteristics as irrelevant, or seek to evaluate them by simulation (which involves a data generation model).

Most Bayesian and frequentist methods for E-CER are based on the comparison of two competing treatments, which sharply pits the statistical efficiency of balanced designs against the ethical mandate to limit exposure to 'losing' treatments. Several recent critiques of adaptive Bayesian or frequentist designs demonstrate that in order to get substantial benefits from adaptive randomization one must pay a high price in operating characteristics, since the efficiency of the comparison between two treatments drops sharply once the imbalance exceeds 2:1 (e.g., Korn and Freidlin (2011)). In contrast, little has been written about the use of adaptive randomization in three (or more) group comparisons, with the notable exception of Berry (2010), who notes '... the adaptive randomization light shines brightest in complicated multiarm settings...'.

In fact, many comparative clinical questions involve more than two competing alternatives. For example, Geriatric Evaluation and Management (GEM) investigators compared four versions of tandem inpatient/outpatient geriatric specialty services for hospitalized older patients (Cohen et al. (2002)). Many behavioral interventions are studied in a standard three-group comparison of usual care, a novel specific intervention, and some kind of 'non-specific' enriched version of usual care. For example, Wilson et al. (BOAT study, 2010) compared 'shared decision-making', 'usual care', and an active control called 'clinician decision-making', which involved guideline-based care but without the specific component of shared decision making with the patient. Biomarker-guided strategy studies of the 'enriched' type may involve comparisons of several standard treatments in multiple strata defined by the biomarkers. Sequential multiple-assignment randomized trials (SMART) often suffer from a 'combinatorial explosion' of individual treatment strategies. Even the comparatively simple two-stage study of rituximab for induction and maintenance in diffuse large B-cell Lymphoma (Habermann et al. (2006); Lunceford, Davidian, and Tsiatis (2002)) has four distinct adaptive treatment strategies to compare, and just adding in options for two rescue treatments increases the number to eight.

Taking these points into consideration, we conjectured that with $K \geq 3$ the benefits of sequential adaptive reallocation of resources away from losing treatments would come without the sharp efficiency penalty associated with adaptive assignment in the K=2 context. We also considered the premise that most of the pairwise treatment effects would be null or nearly so. Finally, we posed the problem as a 'negative selection' task, where pruning unsatisfactory options would be as valuable as declaring superiority of one treatment over another.

2. Treatment Selection

Here we consider the problem of identifying a 'good enough' treatment among available treatments that is not much worse than any other treatment for a prespecified indifference margin. This implies that the selected treatment is not much worse than the unknown best treatment. We propose a group sequential design that modifies the design of Follmann, Proschan, and Geller (1994), that has the goal of selecting the best treatment via monitoring all pairwise comparisons at interim analyses, to the current setting of identifying a 'good enough' treatment. The basic idea, given in more detail below, is to drop a treatment group at interim analysis if there is sufficient evidence that it cannot be much better than the currently observed best treatment.

2.1. Notation

Consider a one-parameter exponential family $f_\theta(x) = e^{\theta x - \psi(\theta)}$ of densities with respect to some measure on the real line. Sufficient statistics for θ are the sample means which are maximum likelihood estimators of $\psi'(\theta)$, and the Kullback-Leibler information number is

$$I(\theta, \lambda) = E_\theta\Big[\log\Big\{\frac{f_\theta(X)}{f_\lambda(X)}\Big\}\Big] = (\theta - \lambda)\psi'(\theta) - \{\psi(\theta) - \psi(\lambda)\}. \qquad (2.1)$$

We assume that for treatment $k = 1, \ldots, K$, independent and identically distributed observations $X_{k1}, X_{k2}, \ldots$ are taken from f_{θ_k}. Let $\theta_{[1]} \leq \cdots \leq \theta_{[K]} = \theta^*$ be the ordered values of $\theta_1, \ldots, \theta_K$, and assume that the treatment with the largest θ is the best treatment. Let δ denote the pre-specified indifference margin between any two treatments.

Consider a group sequential design with $R-1$ interim analyses at sample sizes $N_1, \ldots, N_{R-1}$ and final analysis at sample size N_R. The maximum total sample size N_R is chosen so that a conventional fixed sample size trial that randomizes patients equally to the K treatments and conducts pairwise comparisons of treatments at the end of trial using the Tukey correction (1953), has a desired power (say, 90% power) to detect at least one significant pairwise comparison when the true means are $(\delta, 0, \ldots, 0)$. Let $n_i^{(r)}$ denote the sample size in treatment group i at look r, $r = 1, \ldots, R$. Let $S_i^{(r)} = \sum_{u=1}^{n_i^{(r)}} X_{iu}$ denote the sum of observations in group i at look r, which is the sufficient statistic for θ_i at look r. The maximum likelihood estimate (MLE) of θ_i at look r is $\hat{\theta}_i^{(r)} = S_i^{(r)}/n_i^{(r)}$.

2.2. Non-superiority test

The basic idea of the proposed treatment selection procedure is to drop a treatment at interim analysis if there is sufficient evidence that it cannot be much better than the currently observed best treatment. The determination of ‘sufficient evidence’ can be formulated as a non-superiority test as follows. Let $H_{ij}^{(0)}$ denote the null hypothesis that treatment i is at least δ-better than treatment j for an indifference margin $\delta > 0$, for $i \neq j$:

$$H_{ij}^{(0)} : \theta_i - \theta_j \geq \delta. \tag{2.2}$$

When $H_{ij}^{(0)}$ is rejected, treatment i cannot be δ-better than treatment j (with high confidence) and therefore can be dropped. Suppose there are n_i observations $X_{i1}, \ldots, X_{in_i} \sim f_{\theta_i}$ from treatment group i and n_j observations $X_{j1}, \ldots, X_{jn_j} \sim f_{\theta_j}$ from treatment group j. We can test this hypothesis by using a generalized likelihood ratio test that rejects $H_{ij}^{(0)}$ if

$$\hat{\theta}_i < \hat{\theta}_j + \delta \text{ and } \Lambda_{ij}(\delta) \geq b, \tag{2.3}$$

where $\hat{\theta}_i = \sum_{u=1}^{n_i} X_{iu}/n_i$ and $\hat{\theta}_j = \sum_{v=1}^{n_j} X_{jv}/n_j$ are the MLE of θ_i and θ_j,

$$\Lambda_{ij}(\delta) = \sup_{\theta_i,\theta_j} l_{ij}(\theta_i, \theta_j) - \sup_{\theta_i=\theta_j+\delta} l_{ij}(\theta_i, \theta_j) = n_i I(\hat{\theta}_i, \widetilde{\theta}_i) + n_j I(\hat{\theta}_j, \widetilde{\theta}_j), \tag{2.4}$$

$$l_{ij}(\theta_i, \theta_j) = \sum_{u=1}^{n_i} \log f_{\theta_i}(X_{iu}) + \sum_{v=1}^{n_j} \log f_{\theta_j}(X_{jv}),$$

$\widetilde{\theta}_i$ and $\widetilde{\theta}_j$ are the values of θ_i and θ_j that maximize $l_{ij}(\theta_i, \theta_j)$ under the constraint $\theta_i = \theta_j + \delta$, and the threshold b is chosen according to the given type I error probability using the fact that asymptotically $2\Lambda_{ij}(\delta)$ follows a χ_1^2 distribution under $\theta_i = \theta_j + \delta$.

2.3. Proposed design

Let $I_1 = \{1, \ldots, K\}$ and I_r denote the set of treatment groups that have not been eliminated after the $(r-1)$-th analysis for $r \geq 2$. The initial randomization assigns patients equally among all $K > 2$ treatment groups. After the $(r-1)$-th interim analysis, for $2 \leq r \leq R$, if the study is not stopped at that time, we randomize the next $N_r - N_{r-1}$ patients equally to the treatments in I_r, the treatments that have not been dropped.

The group sequential treatment selection proceeds as follows. At each interim analysis, we drop treatments that (with high confidence) cannot be δ-better than the current observed best treatment. The study is stopped if all treatments other than the currently best treatment are eliminated, and the currently best treatment is selected. After the $(R-1)$-th interim analysis, if two or more treatments remain, the best two treatments are advanced to the last analysis (so $|I_r| = 2$) while other treatments are dropped. At the final analysis, we perform a test of these two treatments with nominal type I error probability. If the test is significant, the treatment with larger θ is selected. If the test is not significant, both treatments are selected.

More specifically, at interim analysis r, $1 \leq r \leq R-1$, let $k_r^* = \arg\max_{i \in I_r} \hat{\theta}_i^{(r)}$ denote the currently observed best treatment. Following (2.3) for testing

$$H_{ik_r^*}^{(0)} : \theta_i - \theta_{k_r^*} \geq \delta, \tag{2.5}$$

we drop treatment group $i \in I_r \backslash \{k_r^*\}$ if

$$\hat{\theta}_i^{(r)} < \hat{\theta}_{k_r^*}^{(r)} + \delta \text{ and } \Lambda_{ik_r^*}^{(r)}(\delta) \geq b, \tag{2.6}$$

where $\Lambda_{ik_r^*}^{(r)}(\delta)$ is the GLR test statistic for the null hypothesis $H_{ik_r^*}^{(0)}$ based on data observed so far. Since, by definition of k_r^*, $\hat{\theta}_i^{(r)} < \hat{\theta}_{k_r^*}^{(r)}$ for every $i \in I_r \backslash \{k_r^*\}$, (2.6) reduces to

$$\Lambda_{ik_r^*}^{(r)}(\delta) \geq b.$$

Note that instead of performing all pairwise comparisons among the remaining treatments in I_r, we compare treatment $i \in I_r \backslash \{k_r^*\}$ to the currently observed best treatment k_r^*. Further, we test the null hypotheses (2.5), $i \in I_r \backslash \{k_r^*\}$, by the reverse order of $\hat{\theta}_i^{(r)}$, starting from the worst treatment (that is, the group with the smallest $\hat{\theta}^{(r)}$), and stopping the pairwise testings when we reach a treatment $j \in I_r \backslash \{k_r^*\}$ that $H_{jk_r^*}^{(0)}$ cannot be rejected. Therefore we perform at most $|I_r| - 1$ pairwise comparisons at interim look r.

The constant b in (2.6) can be chosen such that the probability of a winning treatment that is at least δ better than other treatments has a small chance of being eliminated. For the numerical studies in Section 2.6, we select b using the futility stopping method in Lai and Shih (2004) and Bonferroni correction for $M = K(K-1)/2$ pairwise comparisons. Specifically, for a given $i \neq j$, we choose b such that the probability of rejecting $H_{ij}^{(0)}$ at the $R-1$ interim analyses under $H_{ij}^{(0)}$ is less than or equal to $\epsilon\widetilde{\alpha}/M$ for a small $\epsilon > 0$ (we use $\epsilon = 1/3$ in the examples), where $1-\widetilde{\alpha}$ is the power of the conventional fixed sample size design with the maximum total sample size.

2.4. Relation to other selection procedures

The proposed design is related in its general goal to the sequential multiple comparisons with the best (SMCB) procedures of Paulson (1964), Kao and Lai (1980), and Hsu and Edwards (1983) for selecting the population with the largest mean (or with a 'good enough' mean, within δ of the largest mean) among K normal populations with a common variance. Here we consider the setting of group sequential designs that is more relevant to clinical trials, while the SMCB procedures consider sequential testing in which one observation is added to each remaining group at each stage. The treatment elimination rule and study stopping rule also differ. In SMCB procedures, treatment i is eliminated at stage n when the one-sided test statistic comparing treatment j to treatment i is large for some treatment j that has not yet been eliminated, where 'large' for Paulson's (1964) procedure is derived using bounds on the error probabilities of one-sided sequential probability ratio tests and Bonferroni's corrections to ensure the procedure satisfies the probability of correct selection (PCS) constraint

$$P_{\theta_1,\ldots,\theta_K}\{\theta_D = \theta^*\} \geq P^* \quad \text{whenever } \theta^* \geq \theta_{[K-1]} + \delta \tag{2.7}$$

for a given P^*, where D denotes the selected treatment group, whereas 'large' for the procedures of Kao and Lai (1980) and Hsu and Edwards (1983) is defined to guarantee simultaneous coverage of the confidence intervals at each stage, including when the study is stopped, *for all K treatments versus the best*; both procedures satisfy a stronger PCS constraint

$$P_{\theta_1,\ldots,\theta_K}\{\theta_D > \theta^* - \delta\} \geq P^* \quad \text{for all } \theta_1, \ldots, \theta_K. \tag{2.8}$$

The SMCB procedures stop the study when the population with the largest mean (or a 'good enough' mean) has been found or when it reaches maximum sample size. Our approach eliminates treatments when they are unlikely to be δ-better than the currently observed best treatment. The study is stopped when there is only one treatment left or when it reaches the maximum sample size.

2.5. Nuisance parameters

Nuisance parameters can be easily incorporated into the above procedure. Suppose $\boldsymbol{\theta}_i$ is a $d\times 1$ vector and $\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots$ are i.i.d. $d\times 1$ vectors from treatment group i having density function $f_{\boldsymbol{\theta}_i}(\boldsymbol{x}) = e^{\boldsymbol{\theta}_i^T \boldsymbol{x} - \psi(\boldsymbol{\theta}_i)}$ with respect to a measure ν on $\mathbb{R}^d$. Suppose the first component θ_{i1} of $\boldsymbol{\theta}_i$ is of primary interest, with the corresponding non-superiority null hypothesis

$$H_{ij}^{(0)} : \theta_{i1} - \theta_{j1} \geq \delta.$$

Then the treatment selection procedure can be carried out by replacing the GLR statistic in (2.4) with

$$\Lambda_{ij}(\delta) = \sup_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_j} l_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) - \sup_{\theta_{i1} = \theta_{j1} + \delta} l_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = n_i I(\hat{\boldsymbol{\theta}}_i, \widetilde{\boldsymbol{\theta}}_i) + n_j I(\hat{\boldsymbol{\theta}}_j, \widetilde{\boldsymbol{\theta}}_j),$$

where $\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\theta}}_j$ are the MLE of $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, $\widetilde{\boldsymbol{\theta}}_i$ and $\widetilde{\boldsymbol{\theta}}_j$ are the values of $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ that maximize $l_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ under the constraint $\theta_{i1} = \theta_{j1} + \delta$. One application is to the family of normal distributions with unknown means and variances, which can be expressed as a two-dimensional exponential family for (X, X^2); see Example 1 of Chan and Lai (2000).

2.6. Numerical studies

We conducted two numerical studies to assess the performance of the proposed design for selecting a 'good enough' treatment among $K = 3$ or $K = 4$ treatments. We generated observations from $N(\theta, 1)$ distributions. For normal X with known variance 1, $I(\theta, \lambda) = (\theta - \lambda)^2/2$. It follows that the GLR test statistic for testing $H_{ij}^{(0)}$ at interim look r is

$$\Lambda_{ij}^{(r)}(\delta) = \frac{n_i^{(r)}}{2}(\hat{\theta}_i^{(r)} - \widetilde{\theta}_i)^2 + \frac{n_j^{(r)}}{2}(\hat{\theta}_j^{(r)} - \widetilde{\theta}_j)^2, \tag{2.9}$$

where

$$\hat{\theta}_i^{(r)} = \frac{S_i^{(r)}}{n_i^{(r)}}, \quad \hat{\theta}_j^{(r)} = \frac{S_j^{(r)}}{n_j^{(r)}}, \quad \widetilde{\theta}_i = \frac{S_i^{(r)} + n_j^{(r)}\delta + S_j^{(r)}}{n_i^{(r)} + n_j^{(r)}}, \quad \widetilde{\theta}_j = \frac{S_i^{(r)} - n_i^{(r)}\delta + S_j^{(r)}}{n_i^{(r)} + n_j^{(r)}}.$$

At the end of the trial, one or more treatments may be selected. We describe the selection as full success, partial success, or failure, depending on the true means. When all the treatments that are within δ of the best treatment are selected by the procedure, we call this full success. Partial success refers to the case that each of the selected treatments is within δ of the best treatment, but not all of the treatments within δ of the best treatment are selected. All other cases are considered failures. Note that the probability of correct selection $P_{\theta_1,\ldots,\theta_K}\{\theta_D = \theta^*\}$ considered by Paulson (1964) is the probability of fully correct selection when the best treatment is at least δ-better than the next best treatment, and the probability of correct selection $P_{\theta_1,\ldots,\theta_K}\{\theta_D > \theta^* - \delta\}$ considered by Kao and Lai (1980) and Hsu and Edwards (1983) is one minus the probability of incorrect selection. We also report other important performance measures including the probability of choosing a particular treatment, the probability of selecting exactly r treatments, for $1 \leq r \leq K$, the expected value of the number of the looks T, and the expected final sample size.

The first numerical study considered treatment selection among $K = 3$ treatments with a common $\delta = 0.3$ for all pairwise comparisons, for six scenarios (see also Table 1). In Scenario 1, all treatment means are the same, and any selection of $r < K$ treatments is 'partially correct'; to be fully correct all K treatments must be selected. In Scenario 2 there is a single treatment that is at least δ better than all the other treatments, and only that selection is (fully) correct. In Scenario 3, one treatment is better by $< \delta$ than all the others, and any selection of $r < K$ is (partially) correct, with $r = K$ fully correct, as in Scenario 1. In Scenario 4, there are two treatments both at least δ better than the other treatment but within δ of each other. Selecting one of them is partially correct, all of them, fully correct. In Scenario 5, there is a 'winning' treatment at least δ better than some other 'losing' treatment, but there is one 'middling' treatment within δ of all the others. A selection of some but not all the 'winning and middling' treatments is partially correct, and selecting them all (but excluding the 'losing' treatment) is fully correct. Scenario 6 is similar to Scenario 5 except the 'middling' treatment is closer to the losing treatment.

A fixed sample size design with equal randomization to the $K = 3$ treatments, and using Tukey's method to adjust for all pairwise comparisons with 5% family-wise error rate, requires 164 subjects per arm (total 492 subjects) to have 80% power to detect at least one significant pairwise comparison when the true means are $(0.3, 0, 0)$. Table 1 shows the characteristics of the proposed group sequential design with $R = 3$ looks: two interim looks at sample sizes $N_1 = 165$ and $N_2 = 330$, and final analysis at $N_3 = 492$. Treatments are dropped at interim analyses according to (2.6) with $\delta = 0.3, b = 2.478$. Each result is based on 5,000 simulations. The proposed design has at least 97% probability of being fully or partially correct in selecting a 'winning' treatment. When there are more than one equally winning treatments (Scenarios 1 and 4), these treatments have roughly equal probability of being selected. When there are winning treatments that are at least δ-better than the next best treatment (Scenarios 2 and 4), the probability of correctly selecting one of the winning treatments is at least 97%. Even when the winning treatment is not δ-better than the next best treatment

Table 1. Treatment selection and expected sample size for Simulation Study 1.

Case	True means	Correct selection			Incorrect	Pr(trt k is selected)			Pr(select r trts)			
		full	partial	total	selection	k =1	2	3	r =1	2	$E(T)$	$E(N)$
1	(0.0, 0.0, 0.0)	0.000	1.000	1.000	0.000	0.344	0.334	0.323	0.999	0.001	1.62	267.2
2	(0.3, 0.0, 0.0)	0.970	0.000	0.970	0.030	0.970	0.017	0.013	1.000	0.000	1.26	207.2
3	(0.2, 0.0, 0.0)	0.000	1.000	1.000	0.000	0.873	0.064	0.063	1.000	0.000	1.43	236.5
4	(0.3, 0.3, 0.0)	0.000	0.998	0.998	0.002	0.511	0.487	0.002	1.000	0.000	1.51	249.8
5	(0.3, 0.2, 0.0)	0.000	0.992	0.993	0.007	0.764	0.229	0.007	1.000	0.000	1.46	241.6
6	(0.3, 0.1, 0.0)	0.000	0.989	0.989	0.011	0.918	0.070	0.011	1.000	0.000	1.36	223.7

(Scenarios 3, 5 and 6, with differences ranging from 0.1 to 0.2), the best treatment is correctly selected for over 75% of the times. Compared to the fixed sample size design with sample size 492, the proposed design has substantial savings in expected sample size (ranging from 207.2 to 267.2), with expected number of looks ranging from 1.26 to 1.62.

Table 2(a) shows the results for treatment selection among $K = 4$ treatments with a common $\delta = 0.3$ for all pairwise comparisons. A fixed sample size design with equal randomization to the $K = 4$ treatments, and using Tukey's method to adjust for all pairwise comparisons with 5% family-wise error rate, requires 163 subjects per arm (total 656 subjects) to have 80% power to detect at least one significant pairwise comparison when the true means are $(0.3, 0, 0, 0)$. Table 2(a) shows the characteristics of the proposed group sequential design with $R = 3$ looks: two interim looks at sample sizes $N_1 = 220$ and $N_2 = 440$, and final analysis at $N_3 = 656$. Treatments are dropped at interim analyses using (2.6) with $\delta = 0.3, b = 3.107$. Each result is based on 5,000 simulations. The proposed design has over 96% probability of fully or partially correctly select a 'winning' treatment for all the scenarios considered, with substantial savings in expected sample size (ranging from 275.6 to 433.7) as compared to the conventional fixed sample size design. When there are winning treatments that are within δ of each other and are at least δ-better than the losing treatments without middling treatments (Scenarios 2-7 and 10), the proposed design has over 97% probability of (fully or partially) correctly selecting one of the winning treatments.

Table 2(b) shows the results under a similar setting to Table 2(a) except that the magnitude of the indifference margin depends on the pair of treatments being compared. This is motivated by the GEM study (Cohen et al. (2002)), in which the better treatments are also more expensive. Therefore for cost-effectiveness purposes one may choose the indifference margin between two treatments according to the difference in their cost. To test the null hypothesis that treatment i is at least δ_{ij}-better than treatment j,

$$\widetilde{H}_{ij}^{(0)} : \theta_i - \theta_j \geq \delta_{ij}, \tag{2.10}$$

Table 2. Treatment selection and expected sample size for Simulation Study 2.

(a) Equal indifference margins

Case	True means	Correct selection			Incorrect	Pr(trt k is selected)				Pr(select r trts)			
		full	partial	total	selection	k =1	2	3	4	r =1	2	$E(T)$	$E(N)$
1	(0.0,0.0,0.0,0.0)	0.000	1.000	1.000	0.000	0.285	0.281	0.285	0.289	0.860	0.140	1.97	433.7
2	(0.4,0.0,0.0,0.0)	0.997	0.000	0.997	0.003	1.000	0.001	0.001	0.001	1.000	0.000	1.25	275.6
3	(0.3,0.0,0.0,0.0)	0.969	0.000	0.969	0.031	0.975	0.009	0.014	0.009	0.993	0.007	1.47	324.1
4	(0.4,0.4,0.0,0.0)	0.005	0.994	1.000	0.000	0.496	0.510	0.000	0.000	0.994	0.006	1.64	361.7
5	(0.4,0.3,0.0,0.0)	0.007	0.992	1.000	0.000	0.794	0.213	0.000	0.000	0.993	0.007	1.61	355.1
6	(0.4,0.4,0.4,0.0)	0.000	1.000	1.000	0.000	0.350	0.356	0.344	0.000	0.950	0.050	1.83	401.5
7	(0.4,0.3,0.3,0.0)	0.000	1.000	1.000	0.000	0.684	0.172	0.185	0.000	0.960	0.040	1.78	390.7
8	(0.4,0.3,0.2,0.0)	0.000	1.000	1.000	0.000	0.766	0.220	0.041	0.000	0.972	0.028	1.71	376.0
9	(0.4,0.2,0.2,0.0)	0.000	0.999	0.999	0.001	0.911	0.054	0.054	0.001	0.980	0.020	1.61	355.1
10	(0.4,0.1,0.1,0.0)	0.975	0.000	0.975	0.025	0.979	0.011	0.014	0.001	0.995	0.005	1.42	311.3

(b) Unequal indifference margins

Case	True means	Correct selection			Incorrect	Pr(trt k is selected)				Pr(select r trts)			
		full	partial	total	selection	k =1	2	3	4	r =1	2	$E(T)$	$E(N)$
1	(0.0,0.0,0.0,0.0)	0.000	1.000	1.000	0.000	0.230	0.241	0.303	0.412	0.814	0.186	2.00	439.1
2	(0.4,0.0,0.0,0.0)	0.988	0.000	0.988	0.012	0.991	0.002	0.004	0.007	0.996	0.004	1.38	302.9
3	(0.3,0.0,0.0,0.0)	0.933	0.000	0.933	0.067	0.959	0.013	0.020	0.035	0.973	0.027	1.63	358.0
4	(0.4,0.4,0.0,0.0)	0.006	0.994	0.999	0.001	0.474	0.531	0.000	0.000	0.994	0.006	1.60	351.2
5	(0.4,0.3,0.0,0.0)	0.008	0.990	0.998	0.002	0.774	0.232	0.001	0.001	0.992	0.008	1.60	352.4
6	(0.4,0.4,0.4,0.0)	0.000	1.000	1.000	0.000	0.315	0.343	0.418	0.000	0.923	0.078	1.81	398.9
7	(0.4,0.3,0.3,0.0)	0.000	0.999	0.999	0.001	0.630	0.182	0.264	0.001	0.923	0.077	1.82	399.0
8	(0.4,0.3,0.2,0.0)	0.000	0.999	0.999	0.001	0.733	0.239	0.063	0.001	0.964	0.036	1.72	378.8
9	(0.4,0.2,0.2,0.0)	0.000	0.998	0.998	0.002	0.862	0.069	0.100	0.002	0.967	0.033	1.69	372.0
10	(0.4,0.1,0.1,0.0)	0.961	0.000	0.961	0.039	0.968	0.016	0.018	0.005	0.992	0.008	1.53	336.0

we can use a generalized likelihood ratio test that rejects $\widetilde{H}_{ij}^{(0)}$ if

$$\hat{\theta}_i < \hat{\theta}_j + \delta_{ij} \text{ and } \Lambda_{ij}(\delta_{ij}) \geq b, \tag{2.11}$$

where $\Lambda_{ij}(\cdot)$ is as defined in (2.4). For this simulation study, we assumed that treatments 1-4 are in order of cost (treatment 1 more expensive than treatment 2, and so on); and we set $\delta_{ij} = 0.35$ for $i < j$, $\delta_{21} = \delta_{32} = \delta_{43} = 0.3$, $\delta_{31} = \delta_{42} = 0.25$, $\delta_{41} = 0.2$. Each result of Table 2(b) is based on 5,000 simulations. As expected, when there are multiple treatments with the same mean, the less expensive treatment is selected more often. The proposed design has over 93% probability of fully or partially correctly selecting a 'winning' treatment for the scenarios considered, with large savings in the expected sample size.

3. A Local CER Approach to Guideline Implementation

E-CER has its own peculiar complexities. It usually involves multiple sites (either for accrual or to increase the generalizability of findings) and calls for a relatively permissive implementation to accommodate local site variations in the delivery of care. For example, in extending the original VA POC example (comparing different methods for insulin dosing) from the first site to others, investigators had to deal with the variety of ways that different VA health care centers deliver diabetes care. In one site, the prescribing physician is typically a trainee, working under a chief resident during a short-term rotation, while in another site it is a hospitalist, with a permanent position. Patient populations differ across sites, in race, social class, income, and rate of co-morbid conditions, even within the VA system. The local experience may reflect long experience with one method over another, leading to local variations in 'comfort' with each treatment. Readers of the trials literature often question whether the average effects studied in clinical trials accurately reflect the possible heterogeneity of true effects across subpopulations. For example, clinicians who treat a large number of complex cases (with a great deal of comorbidity) may worry that the trials underlying the guidelines have been done in less challenging patients, a worry that can become a barrier to implementation if it is reinforced with a few bad experiences. That causes some skepticism in local decision-making, where clinicians ask 'Do the results of these trials reflect what will happen if we implement the recommendations here at home?' When that question is not answered satisfactorily, there can be principled opposition to attempts to impose 'evidence-based guidelines' (as opposed to somewhat less well-founded general opposition to new ideas.) Furthermore, many such guidelines are not based on high-quality E-CER, but rather on opinion or observational work, inducing even more informed skepticism. Failure to implement guidelines has been identified as a major impediment to improving cost-effectiveness of care in the US and elsewhere (which in turn is a critical and urgent need to avoid bankruptcy of even the most advanced nations).

One way to recognize and deal with such concerns is to think of an attempt at guideline implementation as an example of local CER. When guidelines are implemented in large systems of providers (such as the VA or other networks) it may prove helpful to perform a randomized CER study *as a form of implementation.* For example, sites subject to implementation would be queried about their possible preferred alternatives to the proposed guideline, and the results assembled into $K > 2$ possible treatments with some existing support in the collection of sites. Then the K treatments would be ranked according to the number of sites that have the treatment on its preferred list (say, the 'unpopularity' score, ranging from 0, all sites prefer a treatment, to 1, no site prefers the treatment). By using the setup in the previous section, replacing cost by unpopularity score, one would assure sites that dropping a popular treatment requires stronger evidence than dropping an unpopular treatment. Furthermore, if the guideline treatment is truly a sufficient improvement over alternatives, implementation is almost automatic in the course of the trial, in the sense that losing alternatives are shed along the way. Of course, using a trial as a locally acceptable method of implementation of a guideline requires the kind of infrastructure that is being developed in the POC-CT project, because it will enable sites to quickly mount a joint multi-site trial without having to go through the funding and approval hoops of a typical CER trial done under the aegis of a research organization, such as the NIH.

4. A Short-cut for Survival Studies with an Early Response Indicator

In other work we have proposed a seamless Phase II-III design (Lai, Lavori, and Shih (2012)) for exploiting an early response indicator that is a sensitive indicator for the effect of treatment on a time to clinical event (such as survival). That is, it is accepted that two treatments that do not differ on an early response rate will not differ appreciably in longer-term definitive outcomes (but not necessarily conversely). The connection to the current work is that one might imagine conducting a trial such as we propose here, with the binary outcome (response) used during the interim looks, because it would allow investigators to discard treatments that have no chance of being contenders on the basis of poor response. Then the surviving treatments at the end would be continued to study the time-to-event difference. Such a design takes advantage of the timeliness of the response outcome, and may be a way to circumvent the usual problem of sequential survival studies that, unless accrual is slow relative to mortality, there is little chance of stopping early for futility or effect. Furthermore, the method anticipates the development of increasingly sensitive early markers of failure of treatment, and could be adapted to handle the large number of treatment policies that result from a two-stage (or more) adaptive treatment strategy trial such as described by Lunceford, Davidian, and Tsiatis (2002), Wahed and Tsiatis (2004), and Ko and Wahed (2012).

5. Discussion

Most discussions of clinical trial designs for CER emphasize the ‘research’ aim, which is to create new knowledge by showing that one treatment is superior to others. This aim leads directly to the traditional significance testing framework, because the assertion of superiority requires rejection of a null hypothesis. But the practical goal of CER is to support evidence-based choices of treatment. That goal can be approached by several paths, including some that do not require statistical proof of superiority. The popular imagination is seized of the notion of a ‘breakthrough treatment’ that dominates its competitors in head-to-head trials, leading to steadily advancing success. But reality is messier, with many choices of treatments in wide use for a given condition, most of them unsupported by reliable evidence, with new treatments being marketed without the need for demonstrations of superiority to existing treatments. Under the circumstances, the best that can be done may be to whittle away the worst choices, leaving a generally higher quality of otherwise indistinguishable competitors.

In contrast to the 'generalizable knowledge' aspect of research, which is intended to apply across health care organizations, regions, and even continents, implementation of such knowledge is inherently local, requiring the participation of individual clinicians and the infrastructure that supports their work. Because of the considerable independence of individual clinicians and the flexibility they are allowed in their practices, the spread of scientific conclusions into practice is by no means automatic. This motivates some novel approaches to CER, such as POC-CT, that integrate the evidence-gathering and implementation phases so that the same clinicians who must implement the results are the generators of the data in the first place. Such approaches are in stark contrast to the standard CER model, in which a professional class of researchers pluck subjects from the usual care stream and conduct highly structured studies designed to produce knowledge that will be handed off for dissemination and implementation by others.

In this paper we explore some possible alternatives to the standard designs, that may facilitate novel approaches to CER. By removing the test of a null hypothesis from its central role in a conventional trial, we can realize considerable reductions in sample size and therefore cost and time, while preserving other desirable operating characteristics that may be better suited to the goals of pragmatic CER.

As one referee pointed out, the scope of the model for our approach can be broadened to semiparametric models such as linear regression with unspecified error distributions, logrank tests, and Cox proportional hazards models, as long as the sequential test statistics, after appropriate normalization, are asymptotically multivariate normal with independent increments. Caution should be made, however, in using the proposed approach for treatment selection based on time to event endpoints. At the interim analysis one only observes the survival curves up to a certain timepoint and therefore could miss late treatment benefit or late treatment harm. Also, unless the events occur relatively quickly as compared to accrual, the chance to detect non-superiority at interim analyses may be small and thus the proposed elimination scheme may not be useful to drop treatments and allocate more patients to the remaining treatment groups. For such scenarios, the procedure described in Section 4 may be more appropriate; it uses an early response indicator to select treatments to be compared on time to events, when there is an early response indicator that is a sensitive indicator for the effect of treatment on the time to event.

One area of future research is the quantification of probabilities of correct selection such as those in (2.7)−(2.8). These probabilities clearly depend on the maximum sample size, the number and timing of interim analyses, and the treatment elimination and selection scheme. In our numerical examples the maximum sample size N_R was chosen such that a fixed sample size trial with equal randomization has 80% power to detect at least one significant pairwise comparison when the true parameters are $(\delta, 0, \ldots, 0)$. We suspect this is the maximum sample size that would be feasible in practice. For simplicity we assumed two interim analyses and one final analysis with roughly equal group sizes. In practice, one may choose the number and timing of interim analyses as in conventional group sequential clinical trials, where multiple factors contribute to the selection of group sizes at interim analyses, such as accrual rate, rate of accumulated information, schedule of DMC meetings, and time needed to clean the data and conduct interim analysis. This is typically determined on a case-by-case basis according to study specifics. Adaptive choice of the group sizes can be considered as well.

In contrast to the SMCB procedures of Paulson (1964), Kao and Lai (1980), and Hsu and Edwards (1983), in which there is no upper bound on the sample size and one can choose the stopping rules to achieve a desired lower bound on the specific probability of correct selection, our procedure with a given maximum sample size and given group sizes at interim analysis does not have this property. However, simulations show that the probabilities of correct selection are greater than 95% for the parameter configurations considered in the numerical examples. We note that the PCS parameter configuration (2.7) may not be particularly plausible or relevant in the E-CER setting, where there may well be two or more ‘good enough’ treatments and one or more inferior treatments, and the salient goal is to deselect the latter, as quickly as possible. As in the conventional setting, it will require developmental work and practical experience to understand how the context of the trial defines the target operating characteristics for the design.

Acknowledgement

This research was supported in part by National Institutes of Health grant 1 P30 CA124435-01 (P. W. Lavori), by the Clinical and Translational Science Award 1UL1 RR025744 for the Stanford Center for Clinical and Translational Education and Research (Spectrum) from the National Center for Research Resources, National Institutes of Health, and by the U.S. Department of Veterans Affairs Cooperative Studies Program (M.-C. Shih).

References

Berry, D. A. (2010). Adaptive clinical trials: the promise and the caution. *J. Clin. Oncol.* **29**, 606-609.

Berry, D. A. (2012). Bayesian approaches for comparative effectiveness research. *Clinical Trials* **9**, 37-47.

Berry, S. M., Carlin, B. P., Lee, J. J. and Müller, P. (2011). *Bayesian Adaptive Methods for Clinical Trials*. Chapman and Hall/CRC Press: Boca Raton, FL.

Chan, H. P. and Lai, T. L. (2000). Asymptotic approximations for error probabilities of sequential or fixed sample size tests in exponential families. *Ann. Statist.* **28**, 1638-1669.

Cohen, H. J., Feussner, J. R., Weinberger, M., Carnes, M., Hamdy, R. C., Hsieh, F., Phibbs, C. and Lavori, P. (2002). A controlled trial of inpatient and outpatient geriatric evaluation and management. *N. Engl. J. Med.* **346**, 905-912.

Fiore, L. D., Brophy, M., Ferguson, R. E., D'Avolio, L., Hermos, J. A., Lew, R. A., Doros, G., Conrad, C. H., O'Neil, J. A. Jr, Sabin, T. P., Kaufman, J., Swartz, S. L., Lawler, E., Liang, M. H., Gaziano, J. M. and Lavori, P. W. (2011). A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen. *Clinical Trials* **8**, 183-195.

Follmann, D. A., Proschan, M. A. and Geller, N. L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* **50**, 325-336.

Habermann, T. M., Weller, E. A. Morrison, V. A., Gascoyne, R. D., Cassileth, P. A., Cohn, J. B., Shaker, R. D. , Woda, B., Fisher, R. I., Peterson, B. A. and Horning, S. J. (2006). Rituximab-CHOP versus CHOP alone or with maintenance rituximab in older patients with diffuse large B-cell lymphoma. *J. Clin. Oncol.* **24**, 3121-3127.

Hsu, J. C. and Edwards, D. G. (1983). Sequential multiple comparisons with the best. *J. Amer. Statist. Assoc.* **78**, 958-964.

Huang, X., Ning, J., Li, Y., Estay, E., Issa, J.-P. and Berry, D. A. (2009). Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Statist. Med.* **28**, 1680-1689.

Kao, S. C. and Lai, T. L. (1980). Sequential selection procedures based on confidence sequences for normal populations. *Comm. Statist. Theory Methods* **9**, 1657-1676.

Ko, J. H. and Wahed, A. S. (2012). Up-front versus sequential randomizations for inference on adaptive treatment strategies. *Statist. Med.* **31**, 812-830.

Korn, E. L. and Freidlin, B. (2011). Outcome-adaptive randomization: Is it useful? *J. Clin. Oncol.* **29**, 771-776.

Lai, T. L. and Shih, M.-C. (2004). Power, sample size and adaptation considerations in the design of group sequential clinical trials. *Biometrika* **91**, 507-528.

Lai, T. L., Lavori, P. W. and Shih, M.-C. (2012). Sequential design of phase II-III cancer trials. *Statist. Med.* **31**, 1944-1960.

Lunceford, J. K., Davidian, M. and Tsiatis, A. A. (2002). Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics* **58**, 48-57.

Paulson, E. (1964). A sequential procedure for selecting the population with the largest mean from k normal populations. *Ann. Math. Statist.* **35**, 174-180.

Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688-701.

Rubin D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6**, 34-58.

Stukel, T. A., Fisher, E. S., Wennberg, D. E., Alter, D. A., Gottlieb, D. J. and Vemeulen, M. J. (2007). Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management of AMI survival using propensity score and instrumental variable methods. *J. Amer. Med. Assoc.* **297**, 278-285.

Tukey, J. W. (1953). *The Problem of Multiple Comparisons.* Mimeographed monograph.

Wahed, A. S. and Tsiatis, A. A. (2004). Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics* **60**, 124-33.

Wilson, S. R., Strub, P., Buist, A. S., Knowles, S. B., Lavori, P. W., Lapidus, J., Vollmer, W. M. and the Better Outcomes of Asthma Treatment (BOAT) Study Group. (2010). Shared treatment decision making improves adherence and outcomes in poorly controlled asthma. *Amer. J. Respir. Crit. Care Med.* **181**, 566-577.

Department of Health Research and Policy, Stanford University, 259 Campus Drive, Stanford, CA 94305, USA.

E-mail: meichiun@stanford.edu

Department of Health Research and Policy, Stanford University, 259 Campus Drive, Stanford, CA 94305, USA.

E-mail: lavori@stanford.edu

(Received July 2012; accepted April 2013)