Journal of Clinical Epidemiology 66 (2013) S130−S137

Journal of Clinical Epidemiology

Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus

Jason T. Connor[a,b,*], Jordan J. Elm[c], Kristine R. Broglio[a], and for the ESETT and ADAPT-IT Investigators

[a]Berry Consultants, 4301 Westbank Dr, Suite 140, Bldg B, Austin, TX 78746, USA
[b]University of Central Florida College of Medicine, 6850 Lake Nona Blvd, Orlando, FL 32827, USA
[c]Department of Public Health Sciences, Medical University of South Carolina, 135 Cannon St, Suite 303, Charleston, SC 29425, USA

Accepted 19 February 2013

Abstract

Objective: We present a novel Bayesian adaptive comparative effectiveness trial comparing three treatments for status epilepticus that uses adaptive randomization with potential early stopping.

Study Design and Setting: The trial will enroll 720 unique patients in emergency departments and uses a Bayesian adaptive design.

Results: The trial design is compared to a trial without adaptive randomization and produces an efficient trial in which a higher proportion of patients are likely to be randomized to the most effective treatment arm while generally using fewer total patients and offers higher power than an analogous trial with fixed randomization when identifying a superior treatment.

Conclusion: When one treatment is superior to the other two, the trial design provides better patient care, higher power, and a lower expected sample size.

Keywords: Comparative effectiveness research; Bayesian adaptive trials; Response adaptive randomization; Adaptive sample size; Status epilepticus; Emergency medicine

1. Background

Regulatory clinical trials are frequently long and expensive and fail to find a statistically significant difference between treatments [1]. Comparative effectiveness trials, those comparing commercially available products, can be even more costly and even less likely to produce statistically significant results, because one is typically searching for a very small effect size in a setting with greater variability. The smaller effect size is because the difference between two effective therapies is likely smaller than that between an effective therapy and a placebo. Variability is often greater if the trial is conducted in a pragmatic setting, a trial with more representative ‘‘real-world’’ inclusion/exclusion criteria, as is encouraged for maximum generalizability [2,3]. This combination requires a much higher sample size for adequate power, producing a longer and/or (usually and) more expensive trial.

When comparing multiple commercially available and perhaps U.S. Food and Drug Administration (FDA)- or European Medicines Agency (EMEA)-approved products or treatment strategies against one another via comparative effectiveness research (CER) trial, Bayesian adaptive trials may offer strong benefit for their ability to calculate the probability that each treatment is the best or worst. Here we describe the design of a multiarm Bayesian adaptive comparative effectiveness trial in refractory status epilepticus that compares three commonly used drugs. The Established Status Epilepticus Treatment Trial (ESETT) is a phase 3 comparative effectiveness trial in patients with established status epilepticus who have failed benzodiazepines.

The goal of this article was to illustrate via a detailed example of an actual trial design, how Bayesian adaptive designs are well suited to comparative effectiveness trials. We discuss the ESETT and how its Bayesian adaptive design is specifically tailored to answer the key clinical question of which treatment to choose by describing the design features, an example of its execution, and its characteristics compared with a more traditional design. The design illustrated here typically offers a lower expected sample size with higher power than a standard fixed-allocation three-arm trial, while

Conflict of interest: The authors have no conflicts of interest related to the content of this article.

* Corresponding author. Tel.: 317-877-1084.

E-mail address: jason@berryconsultants.com (J.T. Connor).

0895-4356/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jclinepi.2013.02.015

What is new?

There are very few examples of applying the novel approaches of Bayesian adaptive trial design in a comparative effectiveness setting. We describe a Bayesian adaptive trial design that uses adaptive randomization in a multi-arm trial to produce higher power, lower total trial size, and a trial with more patients receiving the most efficacious therapies.

also having a high probability of randomizing a higher proportion of patients to the most effective therapy.

2. Design

This trial was designed as one of five trials emanating from the Adaptive Designs Accelerating Promising Trials into Treatments (ADAPT-IT) project, a collaborative effort supported by both the National Institutes of Health (NIH) and FDA to explore how adaptive clinical trial design might improve the evaluation of drugs and medical devices [4].

The primary objective of ESETT was to identify the most effective and/or the least effective treatment among three commonly used second-line therapies for status epilepticus within an emergency department setting. There were three treatment arms: fosphenytoin (fPHT), levetiracetam (LVT), and valproic acid (VPA). We define a positive response to treatment as achieving the primary endpoint of clinical cessation of status epilepticus within 20 minutes of the start of study drug infusion without recurrent seizures, life-threatening hypotension, or cardiac arrhythmia within 1 hour.

A maximum sample size of 720 unique patients will provide high (90%) power to identify a single arm as best if it has a response rate at least 15% greater than that of the others. The primary analysis is an intent-to-treat analysis and includes all patients as they were randomized regardless of the treatment actually received or missing outcome. Re-enrollment of the same subject is expected to occur 5% of the time (and only the data associated with the first enrollment will be analyzed). Given the possibility of re-enrollers, treatment crossovers, and missing data, the maximum sample size was inflated from 720 to 795. Thus, the actual trial will enroll a minimum of 400 and a maximum of 795 patients. We expect the primary analysis to include approximately 720 eligible patients.

The trial has predefined adaptations in which randomization probabilities are updated to increase the proportion of patients randomized to superior treatments and to increase study power when a most effective therapy exists [5,6]. The trial also includes interim analyses for stopping early for success or futility if it becomes evident that differences between drugs are unlikely to be identified.

2.1. Statistical model

Each of the three treatment arms is modeled independently. We assume the probability of response, θ_T, has a uniform Beta prior distribution

$$[\theta_\mathrm{T}]\sim\mathrm{Beta}(1,1) \quad \text{for } T\in\{\mathrm{fPHT},\mathrm{LVT},\mathrm{VPA}\}.$$

This is the standard reference prior when estimating a proportion, is conjugate to the binomial distribution, and assigns equal prior probability to all possibilities of the unknown response rates [7]. It is equivalent to starting with two patients' worth of information, one a treatment success and the other a treatment failure.

At each interim analysis, the number of observed responses on each treatment, X_T, among the currently enrolled patients on that treatment, N_T, follows a binomial distribution; therefore, the posterior distribution for each response rate is

$$[\theta_T|X_T,N_T]\sim\mathrm{Beta}(1+X_T,1+N_T-X_T).$$

The treatment with the highest true (but unknown) response rate is labeled $t_{\max}$, whereas the treatment with the lowest response rate is labeled $t_{\min}$. During the trial, we will not know which treatment is $t_{\max}$ and which is $t_{\min}$; however, we can calculate probabilities that each of the three treatments is $t_{\max}$ and $t_{\min}$.

The probability that treatment T is the most effective treatment is expressed as $\Pr(T = t_{\max}) = \Pr(\theta_T > \theta_x$ and $\theta_T > \theta_Y)$, where X and Y represent the two treatments other than treatment T.

The probability that each treatment arm offers the highest response rate can be shown (using LVT as an example) as

$$\Pr(\mathrm{LVT}=t_{\max})=\int_0^1\int_0^{\theta_{\mathrm{LVT}}}\int_0^{\theta_{\mathrm{LVT}}} f(\theta_{\mathrm{LVT}}|X_{\mathrm{LVT}},N_{\mathrm{LVT}})f(\theta_{\mathrm{VPA}}|X_{\mathrm{VPA}},N_{\mathrm{VPA}})f\left(\theta_{\mathrm{fPHT}}\middle|X_{\mathrm{fPHT}},N_{\mathrm{fPHT}}\right)\mathrm{d}\theta_{\mathrm{fPHT}}\mathrm{d}\theta_{\mathrm{VPA}}\mathrm{d}\theta_{\mathrm{LVT}}$$

Likewise, the probability that treatment T is the least effective treatment is expressed as $\Pr(T = t_{\min}) = \Pr(\theta_T < \theta_x$ and $\theta_T < \theta_Y)$, where X and Y represent the two treatments that are not treatment T.

2.2. Primary efficacy analysis

At the conclusion of the trial, we will report the response rate for each treatment group with 95% credible intervals as well as the pair-wise differences in response rates with 95% credible intervals. We will also report the probability that each therapy offers the best and worst response rates, $\Pr(T = t_{\max})$ and $\Pr(T = t_{\min})$.

Currently, all three drugs are commonly used for status epilepticus. Ordering the drugs by effectiveness is the primary goal. Therefore, identifying a treatment as either superior so it can be more broadly used or inferior so its use can be limited will be clinically beneficial. Therefore, this trial will be considered a success if it identifies the most effective treatment or least effective treatment with high probability

$$\Pr(T = t_{\max}) > 0.975 \text{ or } \Pr(T = t_{\min}) = 0.975$$

for a treatment T.

2.3. Adaptive allocation

Initially, 100 patients will be allocated to each arm. After the initial 300 patients, adaptive randomization will begin. Adaptive randomization will focus on identifying the treatment arm offering the highest response rate, labeled $t_{\max}$, using information weighting. Information is a measure of the expected reduction in variance from adding an additional patient to treatment arm T and is defined for treatment arm, T, as

$$I_T = \sqrt{\frac{\Pr(T = t_{\max})\mathrm{Var}(\theta_T)}{N_T + 1}}$$

I_T is calculated for all three treatment arms, and the values are rescaled to produce randomization probabilities $r_T = I_T / \sum(I_t)$ that sum to 1. Therefore, the randomization probability to arm T is proportional to the probability that the arm offers the highest response rate, $\Pr(T = t_{\max})$, and the variance of the response rate estimate, $\mathrm{Var}(\theta_T)$, and inversely proportional to the sample size, N_T. The result is that better treatments are favored, but if at an interim analysis two arms are equally effective, the arm with fewer patients randomized to it will have a larger randomization probability for the next set of patients.

If the adaptive randomization probability for an arm is less than 5%, then that arm is suspended and the remaining arms receive proportionally increased probability. Adaptive randomization probabilities will be updated after every 100 patients are enrolled.

Allocation to the treatment arms will be stratified by age group (2–16, 17–65, and >65 years). Simple blocking within age group is not possible with adaptive randomization probabilities. To ensure similar randomization probabilities across the three age groups while incorporating adaptive randomization, we will use a "Step Forward" centralized randomization procedure developed for emergency treatment trials, as described by Zhao et al. [8].

The timing of analyses, starting at 300 patients and repeating after every 100 patients are enrolled, was chosen after comparing numerous alternative design options via simulation as well as considering the logistical challenges of more frequent randomization updates.

2.4. Interim monitoring for success

Interim monitoring for success will begin after 400 patients have been enrolled and will be repeated after every additional 100 patients are enrolled. Early success stopping is based on identifying a superior treatment. This trial will stop early for success if we have identified the maximum effective treatment with at least 97.5% probability, that is, if any arm $T \in \{\text{fPHT, LVT, VPA}\}$ offers

$$\Pr(T = t_{\max}) \geq 0.975.$$

2.5. Interim monitoring for futility

Interim monitoring for futility will begin after 400 patients have been enrolled and will be repeated after every additional 100 patients are enrolled. Each arm will be monitored independently and terminated if there is a clinically unacceptable response rate. If the probability that a treatment offers at least a 25% response rate is less than 5%,

$$\Pr(\theta_T \geq 25\%) < 0.05,$$

then that arm will be terminated. If all arms have a clinically unacceptable response rate, the trial will be stopped for futility.

The second futility stopping criterion applies if the trial is unlikely to achieve its primary objective, that is, to identify the most effective and/or the least effective treatment. This trial will stop early for futility if the predictive probability of identifying either the most effective ($t_{\max}$) or the least effective treatment ($t_{\min}$) at the maximum sample size is less than 5% [9].

3. Example trial

Table 1 shows example interim analyses for a hypothetical trial that stops early for success. These data are from one of the thousands of simulated trials that inform the trial operating characteristics.

At the first interim analysis, when 300 patients are enrolled, the response rates are 51% for fPHT, 55% for LVT, and 64% for VPA. There is an 88% chance that VPA truly offers the highest response rate and a 70% chance that fPHT has the lowest response rate. The randomization probabilities for the next 100 patients are calculated to be 12%, 22%, and 66% for fPHT, LVT, and VPA, respectively. Of the next 100 patients enrolled, 11, 26, and 63 patients are randomized to fPHT, LVT, and VPA, respectively.

At the 400-patient interim analysis, the response rates are 51%, 59%, and 64% for fPHT, LVT, and VPA, respectively.

Table 1. Example trial demonstrating data gathered at each interim analysis, probability that each treatment arm offers the highest and lowest response rates, randomization probabilities for the next 100 patients, and the predictive probability of identifying the best or worst treatment at the maximum sample size

	Observed responses/randomized (%)			Probability t_{max} (probability t_{min})			Randomization probabilities for next 100 patients			Predictive probability at maximum
N	**fPHT**	**LVT**	**VPA**	**fPHT**	**LVT**	**VPA**	**fPHT**	**LVT**	**VPA**	
300	51/100 (51)	55/100 (55)	64/100 (64)	0.025 (0.70)	0.092 (0.29)	0.88 (0.014)	0.12	0.22	0.66	0.71
New	6/11 (55)	19/26 (73)	39/63 (62)							
400	57/111 (51)	74/126 (59)	105/163 (64)	0.010 (0.87)	0.16 (0.13)	0.83 (0.008)	0.094	0.34	0.57	0.50
New	5/12 (42)	20/38 (53)	34/50 (68)							
500	62/123 (50)	94/164 (57)	139/213 (65)	0.004 (0.88)	0.056 (0.12)	0.94 (0.002)	0.080	0.23	0.69	0.59
New	3/3 (100)	17/28 (61)	55/69 (80)							
600	65/126 (52)	111/192 (58)	194/282 (69)	0.000 (0.87)	0.008 (0.13)	0.992 (0.00)	—	—	—	—

The observed benefit of VPA over LVT has decreased from 9% to 5% but is based on more data. Because the probability that VPA has the highest response rate decreased slightly from 88% to 83%, the randomization probability to LVT increases further for the next 100 patients. The predictive probability of trial success (the probability of identifying the most or least effective treatment at the maximum sample size) is 0.50, well above the 5% threshold to stop for futility. Therefore, the trial continues to the next interim analysis after 500 patients are enrolled.

VPA has the best response rate in the next 100 patients. The probability that VPA offers the highest response rate is 94%, which is high but does not meet the 97.5% required to stop early for success. The trial continues enrollment.

Of the next 100 patients, we randomize 69% to VPA, 23% to LVT, and 8% to fPHT. At the 600-patient interim analysis, VPA has a 99.2% chance of having the highest response rate. This exceeds the 97.5% criterion, and the trial is stopped early, having identified VPA as the most effective treatment. In addition, fPHT has an 87% chance of being the least effective treatment.

4. Operating characteristics

Closed-form solutions for trial operating characteristics such as power, type I error, sample size distribution, and proportion of patients expected to be randomized to each arm do not exist. Therefore, we calculate these key trial characteristics via simulation. We simulate data from using known ‘‘true’’ response rates for each therapy and execute trials incorporating the adaptations described. For each trial, we track the total sample size and number randomized to each therapy, which drug was identified as the best or worst, and so forth. Repeating this process 1,000 times per scenario, we can estimate the trial operating characteristics.

During the design stage, we compared these operating characteristics among competing designs (e.g., more frequent interim looks, fixed randomization, starting at 200 patients vs. 300 patients).

To evaluate how the design performs, we simulated the trial considering different response scenarios. Operating characteristics are based on 1,000 simulations per scenario. We explored six scenarios that represent a broad range of potential treatment effects. Three scenarios include when the three treatments are equivalent but at varying levels of response: all 50%, 25%, or 10%. Other scenarios explore when one arm is 15% better than the other two, when two arms are equally better (15%) than the third, and one in which the response rates decrease across the arms: 65%, 57.5%, and 50%. The simulation results are presented assuming a maximum sample size of 720 unique patients.

Table 2 shows the probabilities of trial success for each of the six response rate scenarios: identifying the best treatment (early, at the maximum sample size, and overall), the worst treatment, and either the best or the worst.

Table 2. Results of 1,000 simulated trials for each of six response rate scenarios

	Probability identify				
Scenario (response rates)	t_{max} **Early**	t_{max} **Max** ***N***	t_{max}	t_{min}	t_{max} **or** t_{min}
Null (0.50–0.50–0.50)	0.012	0.001	0.013	0.018	0.031
One Good (0.50–0.50–0.65)	0.879	0.013	**0.892**	0.033	0.902
Two Good (0.50–0.65–0.65)	0.115	0.003	0.118	**0.672**	0.763
One Middle, One Good (0.50–0.575–0.65)	0.481	0.022	**0.503**	**0.245**	0.682
All Bad (0.25–0.25–0.25)	0.016	0.001	0.017	0.030	0.044
All Really Bad (0.10–0.10–0.10)	0.006	0.000	0.006	0.000	0.006

t_{max} Early and t_{max} Max *N* show the proportion of trials in which an arm is identified as the best treatment at an interim analysis and at the end of the trial, respectively. t_{max} is the total probability of identifying a superior treatment. t_{min} is the probability of identifying the least effective treatment at the end of the trial. t_{max} or t_{min} is the probability of identifying either (or both). Values are bolded denoting true positives under t_{max} and t_{min}.

The false-positive rate for this trial is the probability of declaring an arm to be either the most or the least effective when in truth there is no difference between the arms. This is illustrated in the three "Null" scenarios in which all treatment arms have the same response rate. The probability of identifying a maximum effective treatment is 0.013, 0.017, and 0.006 for equal treatment effects of 0.50, 0.25, and 0.10, and the probability of identifying a least effective treatment is 0.018, 0.030, and <0.001 for the same three scenarios. Thus, the false-positive rate in the all-0.50 response scenario is 0.031. In the "All Bad" and "All Really Bad" scenarios, the false-positive rates are 0.044 and 0.006, respectively. In the "Two Good" scenario, two of the treatment arms have the same response rate, and the probability of identifying one of these as the $t_{\max}$ is 0.12. Thus, the false-positive rate in this scenario is 0.12. However, the clinical consequences of a type I error in CER may be less than those in placebo-controlled trials, because an effective therapy will still be provided to patients. If two treatments are truly similar, but one is erroneously ruled superior, patients still receive an effective therapy. This form of type I error may be less consequential than a situation in which a treatment is erroneously determined to be better than a placebo and patients pay for and receive an ineffective therapy while also being exposed to its adverse events and costs.

Alternatively, the power (the true-positive rate) of this trial is the probability of identifying either the most or the least effective treatment when there truly are differences in the response rates. In the "One Good" scenario, the probability of identifying a maximum effective treatment is 89%. In the "One Middle, One Good" scenario, the probability of identifying either a maximum or a least effective treatment is 68%.

Table 3 shows the mean sample size and the mean allocation to each of the three treatment arms for each of the response rate scenarios. The most effective treatment arms are shown in bold italics. Adaptive allocation leads to a higher proportion of patients on the most effective treatment arms. When two arms are tied for most effective, any patients randomized to either of these arms are included in the percent randomized to most effective.

Table 4 shows, for each arm, the probability that it will be identified as the maximum effective treatment ($t_{\max}$) and the probability that each arm will be identified as the maximum effective treatment with at least 97.5% probability (i.e., reaches the success criteria). The arms with the highest true response rate are shown in bold italics. In the "One Good" scenario, the arm with the highest true response rate has the highest response rate 99.5% of the time and fulfills the success criteria, $\Pr(\text{VPA} = t_{\max}) \geq 0.975$, 89% of the time. In the "Two Good" scenario, two arms have the same response rate. Thus, the probability of being $t_{\max}$ is split between these two arms, and each receives approximately 50% probability of being $t_{\max}$. In the One Middle, One Good scenario, the arm with the highest response rate is identified as the $t_{\max}$ with 95% of the time, and this arm achieves the success criteria: to be clearly identified as the best 50% of the time.

4.1. Comparison to fixed randomization

Table 5 compares the design described here with a similar trial with fixed randomization.

The fixed trials hold a sample size advantage in the Null and All Bad cases, in which all three treatments offer the same treatment effect. Here trials with adaptive randomization enrolled an average of 8–15 more patients.

When treatment effect differences do exist, Table 5 demonstrates that power is maintained, generally with a lower total sample size, while a higher proportion of patients are randomized to the superior treatment. When one treatment offers a 65% response rate vs. 50% in the other two groups, adaptive randomization offers higher power (90% vs. 88%) and lower mean sample size (483 vs. 497), all while randomizing a higher proportion of patients to the better treatment (48% vs. 33%). The One Middle, One Good case further shows similar power, smaller sample size, and a higher proportion of patients randomized to the superior treatment compared with a trial with fixed randomization.

Power to identify either the best or the worst treatment is lower in the Two Good case, because fixed randomization is better at identifying the worst treatment when one treatment is truly worse than then other two. In the adaptive randomization case, fewer patients (16% vs. 33%) are randomized to the worst treatment, leading to lower power: a 79% chance of identifying the worst with fixed randomization decreases to a 67% chance with adaptive randomization. This was a key point of discussion during the design

Table 3. Average sample size based on 1,000 simulations per response rate scenario (Total) and average patients per treatment arm

Scenario (response rates)	Total	fPHT, *n* (%)	LVT, *n* (%)	VPA, *n* (%)
Null (0.50–0.50–0.50)	507	169 (33)	169 (33)	168 (33)
One Good (0.50–0.50–0.65)	483	126 (26)	127 (26)	***230 (48)***
Two Good (0.50–0.65–0.65)	679	115 (17)	***282 (42)***	***282 (42)***
One Middle, One Good (0.50–0.575–0.65)	586	122 (21)	189 (32)	***275 (47)***
All Bad (0.25–0.25–0.25)	524	173 (33)	172 (33)	179 (34)
All Really Bad (0.10–0.10–0.10)	400	133 (33)	133 (33)	134 (33)

Average randomization rates are shown in parentheses. Bold italics indicate the most effective treatment arms.

Table 4. The average Pr($T = t_{max}$) for each $T \in$ {fPHT, LVT, VPA} based on 1,000 simulations per scenario and the proportion of trials in which Pr($T = t_{max}$) $\geq$ 0.975, which indicates that the best therapy has been clearly identified

	Average t_{max}			**Proportion Pr(t_{max}) > 0.975**		
Scenario (response rates)	**fPHT**	**LVT**	**VPA**	**fPHT**	**LVT**	**VPA**
Null (0.50–0.50–0.50)	0.34	0.34	0.32	0.004	0.005	0.004
One Good (0.50–0.50–0.65)	0.001	0.004	***0.995***	0.00	0.00	***0.89***
Two Good (0.50–0.65–0.65)	0.007	***0.48***	***0.51***	0.00	***0.06***	***0.05***
One Middle, One Good (0.50–0.575–0.65)	0.003	0.04	***0.95***	0.00	0.002	***0.50***
All Bad (0.25–0.25–0.25)	0.30	0.34	0.36	0.003	0.009	0.005
All Really Bad (0.10–0.10–0.10)	0.30	0.35	0.34	0.002	0.001	0.003

The arms with the highest true response rate are shown in bold italics.

stage. The clinical team was asked the importance of increasing the probability of identifying the worst treatment, realizing that increasing the probability meant randomizing more patients to a knowingly inferior treatment and decreasing the probability of identifying a superior treatment if one exists. It was decided that it was better to increase power to identify the best therapy and increase the proportion of patients on the better therapies with the trade-off of decreasing power to identify the worst therapy.

5. Discussion

This Bayesian adaptive trial design for CER more closely mimics the goal of continuous quality improvement than would a traditional fixed design and combines that behavior with a prospectively defined protocol that enables us to calculate operating characteristics such as type I error rate and power.

Most importantly in a case in which one treatment is superior to the other two (e.g., second scenario in Table 5), the trial design presented here offers higher power (90% vs. 88%) with a lower expected sample size than a standard, fixed-randomization rate trial (483 vs. 497), all while randomizing a far greater proportion of patients to the superior treatment (48% vs. 33%). In situations in which the three treatments are equal, this design tends to have a slightly larger (8–15 patients) expected sample size than if we had used fixed randomization. Furthermore, type I errors are small: when all three treatments are equally effective, the probability of erroneously declaring one the best is less than 0.02 and the probability of erroneously declaring one the worst is also less than 0.03.

Another key point is that when randomization probabilities drop to less than 5%, randomization to that arm is suspended, but the arm is not dropped and may re-enter at subsequent interim analyses. Situations in which the most effective arms were even temporarily dropped were extremely rare in the simulation study.

In this trial, the main challenges to adaptive randomization are logistical. The trial is conducted in emergency departments with a waiver of informed consent. To speed treatment to patients, we forego any voice- or Internet-based randomization process. Instead, three boxes containing intravenous study drugs are placed in each site, one for each age stratum, and caregivers are instructed to grab the top box, labeled ‘‘Use Next’’, for use. At each allocation update, sites will be instructed how to reorder boxes (each box will have a unique code) based on the centralized randomization scheme.

Design characteristics were chosen via a tradeoff of logistical practicality and the operating characteristics they produce. For example, updating randomization probabilities more often (e.g., every week or every 20 patients) may slightly improve operating characteristics, but given that boxes will have to be reordered at the time of updates to the allocation, we believed that every 100 patients (or approximately every 6 months) would not be overly burdensome to clinical sites. We also believe that minimizing reordering of drugs would decrease the probability of human error.

By writing tailored simulation software (in R), we studied the design over a range of variants which led to

Table 5. Comparison of the design with adaptive randomization to a trial with fixed randomization via power, mean sample size, and the proportion of patients randomized to the therapy with the highest response rate

	Adaptive randomization			**Fixed randomization**		
Scenario (response rates)	**Power**	**Mean *N***	**% to Best**	**Power**	**Mean *N***	**% to Best**
Null (0.50–0.50–0.50)	0.031	507	N/A	0.029	499	N/A
One Good (0.50–0.50–0.65)	0.90	483	48	0.88	497	33
Two Good (0.50–0.65–0.65)	0.76	679	84	0.86	687	67
One Middle, One Good (0.50–0.575–0.65)	0.68	586	47	0.69	599	33
All Bad (0.25–0.25–0.25)	0.044	524	N/A	0.030	509	N/A
All Really Bad (0.10–0.10–0.10)	0.006	400	N/A	0.028	400	N/A

Abbreviations: % to Best, average proportion of patients randomized to the most effective therapy; Power, probability of identifying a treatment as best or worst at the 0.975 level; N/A, not applicable.

The same early stopping rules are used in both. All values are based on 1,000 simulations per scenario.

choosing optimal design features. For example, we discovered the lag between beginning adaptive randomization after 300 patients and the possibility of early stopping after 400 patients helps to decrease the type I error rate compared with starting both at 300 patients. When all three treatments are truly equal, but at the 300-patient analysis one treatment is doing much better because of random variation, this design increases the randomization probability to the best-performing treatment in the next stage. If $\Pr(T = t_{\max}) = 0.975$, a situation in which the trial may stop if early stopping were allowed at 300 patients, instead approximately 82 of the next 100 patients will be randomized to that drug. During that time, we are likely to see the effect size regress to its true mean and not meet the early stopping threshold at the 400-patient interim analysis. However, if the effect were real, most patients are randomized to the best therapy while we confirm it as superior.

The design required custom-written R [10] code (available from the authors) rather than using off-the-shelf sample size software. We and others [11,12] advocate using simulation for trial design even in more standard trials, for example, group sequential designs, as it allows for the illustration of example trials to physician-collaborators, institutional review boards, and data monitoring committees, as well as for the designers to better understand how each adaptive component affects the overall design properties.

The Bayesian paradigm may offer another key advantage in CER. Effect size differences may be small, leading to nonstatistically significant differences in effect sizes. Although a *P*-value cannot be interpreted as a measure of effect size, a clinician still needs to decide which treatment to use. In a frequentist trial, the clinician is left to compare point estimates of the nonsignificantly different therapies. Bayesian posterior probabilities, even if one is not dramatically high, can offer insight to clinicians on the likelihood that they are using the best therapy.

Trial designs should be situation dependent and tailored to the primary clinical objectives. One key consideration in implementing adaptive randomization, in particular, is the time elapsed from when a patient is randomized until he reaches his final endpoints. Here it is a matter of hours or days; so it is straightforward to use accumulating data to influence future randomization probabilities or to stop the trial early for success or futility. In situations with rapid accrual and/or long-term endpoints, such response adaptive randomization may not be feasible.

Adaptive randomization is oftentimes criticized because drift in the probability of response because of changing patient populations over time may lead to bias in parameter estimates. Cook [13] showed, however, that these effects are generally very small. When adaptive randomization is being used, we would caution against dramatic changes in inclusion/exclusion criteria that may lead to changes in overall response rates, or adding additional high-volume sites that may be quite different than the existing sites.

Finally, an adaptive trial requires a system to manage study drug inventory at the site (e.g., reordering of study drug IV bags), which is tied to the randomization of subjects and keeps the clinical sites blinded and the central pharmacy unblinded. When choosing to implement an adaptive design, one needs high confidence that such interactions are likely to produce a trial run according to protocol. Here we plan to use two existing trial networks that have strong and successful experience conducting high-quality trials: the Neurological Emergencies Treatment Trials (NETT) Network with 17 preexisting study sites, Clinical Coordination Center (at the University of Michigan), and Statistical and Data Management Center (at the Medical University of South Carolina) [14,15], and the Pediatric Emergency Care Applied Research Network for pediatric patients [16,17].

6. Summary

We illustrate that Bayesian adaptive trial designs are particularly well suited to comparative effectiveness trials. Randomization probabilities may be updated during the course of the trial to improve patient outcomes while, in a more-than-two-arm study, increasing study power. Furthermore, early stopping may be incorporated so that clinically important results can be shared with the broader community as soon as they are established and confirmed within the prospective trial setting.

Acknowledgments

The ADAPT-IT project is supported jointly by the NIH Common Fund and the FDA, with funding administered by the National Institutes of Neurological Disorders and Stroke (NINDS) (U01NS073476). The NETT Network Clinical Coordinating Center (U01NS056975) and Statistical and Data Management Center (U01NS059041) are funded by the NINDS.

The authors have no financial stake in any of the three companies making the three drugs involved in this trial, nor have any of the three drugmakers had any input into the trial design or analysis plan.

References

[1] Luce BR, Kramer JM, Goodman SN, Connor JT, Tunis S, Whicher D, et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. Ann Intern Med 2009;151:206−9.

[2] Chalkidou K, Tunis S, Whicher D, Fowler R, Zwarenstein M. The role for pragmatic randomized controlled trials (pRCTs) in comparative effectiveness research. Clin Trials 2012;9:436−46.

[3] Sox HC, Goodman SN. The methods of comparative effectiveness research. Annu Rev Public Health 2012;33:425−45.

[4] Meurer WJ, Lewis RJ, Tagle D, Fetters MD, Legocki L, Berry S, et al. An overview of the Adaptive Designs Accelerating Promising Trials into Treatments (ADAPT-IT) project. Ann Emerg Med 2012;60:451−7.

[5] Berry DA. Adaptive clinical trials: the promise and the caution. J Clin Oncol 2011;29:606−9.

[6] Meurer WJ, Lewis RJ, Berry DA. Adaptive clinical trials: a partial remedy for the therapeutic misconception. JAMA 2012;307:2377−8.

[7] Berry DA, Stengl DK. Bayesian biostatistics. New York: CRC Press; 1996.

[8] Zhao W, Ciolino J, Palesch Y. Step-forward randomization in multicenter emergency treatment clinical trials. Acad Emerg Med 2010;6: 659−65.

[9] Berry SM, Carlin BP, Lee JJ, Muller P. Bayesian adaptive methods for clinical trials. Chapter 5. Boca Raton: CRC Press; 2011.

[10] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.

[11] Gaydos BG, Anderson KM, Berry D, Burnham N, Chuang-Stein C, Dudinak J, et al. Good practices for adaptive clinical trials in pharmaceutical product development. Drug Inf J 2009;43:539−56.

[12] Quinlan J, Gaydos B, Maca J, Krams M. Barriers and opportunities for implementation of adaptive designs in pharmaceutical product development. Clin Trials 2010;7:167−73.

[13] Cook, JD. The effect of population drift on adaptively randomized trials. UT MD Anderson Cancer Center Department of Biostatistics Working Paper Series. Working Paper 39. 2007.

[14] Silbergleit R, Durkalski V, Lowenstein D, Conwit R, Pancioli A, Palesch Y, et al, NETT Investigators. Intramuscular versus intravenous therapy for prehospital status epilepticus. N Engl J Med 2012; 366:591−600.

[15] Hill MD, Martin RH, Palesch YY, Tamariz D, Waldman BD, Ryckborst KJ, et al, on behalf of the ALIAS Investigators and the Neurological Emergencies Treatment Trials (NETT) Network. The albumin in acute stroke part 1 trial. Stroke 2011;42:1621−5.

[16] Chamberlain JM, Capparelli EV, Brown KM, Vance CW, Lillis K, Mahajan P, et al. Pharmacokinetics of intravenous lorazepam in pediatric patients with and without status epilepticus. J Pediatr 2012;160: 667−72.

[17] Alessandrini EA, Alpern ER, Chamberlain JM, Shea JA, Holubkov R, Gorelick MH, PECARN. for the Developing a diagnosis-based severity classification system for use in emergency medical services for children. Acad Emerg Med 2012;19:70−8.