Published in final edited form as:

J Biopharm Stat. 2017 ; 27(4): 659–672. doi:10.1080/10543406.2016.1198367.

A Novel Evaluation of Optimality for Randomized Controlled Trials

Jo Wick¹, Scott M. Berry^{1,2}, Hung-Wen Yeh^{1,5}, Won Choi^{5,6}, Christina M. Pacheco^{4,5}, Christine Daley^{4,5,6}, and Byron J. Gajewski^{1,3,4,5}

¹Department of Biostatistics, The University of Kansas Medical Center, Kansas City, KS 66160

²Berry Consultants, 4301 Westbank Drive, Suite 140, Bldg B, Austin, TX 78746

³School of Nursing, The University of Kansas Medical Center, Kansas City, KS 66160

⁴Department of Family Medicine, The University of Kansas Medical Center, Kansas City, KS 66160

⁵Center for American Indian Community Health, The University of Kansas Medical Center, Kansas City, KS 66160

⁶Department of Preventative Medicine and Public Health, The University of Kansas Medical Center, Kansas City, KS 66160

SUMMARY

Balanced two-arm designs are more powerful than unbalanced designs and, consequently, Bayesian adaptive designs (BAD) are less powerful. However, when considering other subject- or community-focused design characteristics, fixed two-arm designs can be suboptimal. We use a novel approach to identify the best two-arm study design, taking into consideration both the statistical perspective and the community's perception. Data Envelopment Analysis (DEA) was used to estimate the relative performance of competing designs in the presence of multiple optimality criteria. The two-arm fixed design has enough deficiencies in subject- and community-specific benefit to make it the least favorable study design.

Keywords

Longitudinal; accrual; community based participatory research; American Indians

1. INTRODUCTION

Traditional evaluation of clinical trial designs revolves around efficiency—the optimal design is the one for which a goal can be achieved using the fewest subjects. With the rapid rise in popularity of adaptive designs (e.g., Jennison & Turnbull, 2000), other criteria have been added to the decision-making process. For example, adaptive seamless phase II/III designs have been shown to reduce the time-to-conclusion while maintaining traditional

* Address for correspondence: The University of Kansas Medical Center, School of Medicine, Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS 66160, Phone: 913-588-1603, Fax: 913-588-0252, bgajewski@kumc.edu.

frequentist power and type I error rates (e.g., Inoue, et al., 2002; Kelly, et al., 2005). Response-adaptive designs, in general, are attractive because in many cases they can reduce the proportion of subjects randomized to unsafe and/or ineffective treatments (e.g. Berry et al., 2010; Berry et al., 2011). Practical and ethical considerations like these play an even bigger role in the selection of the optimal design when the research being done is targeting historically underserved or alienated populations. In American Indians (AI), for example, mistrust of investigators is a known deterrent to participation in randomized trials (National Medical Association, 2000). This mistrust is grounded in documented examples of unethical medical research. In fact, studies that have directly examined the attitudes of ethnic minorities have shown that fear and mistrust of medical research deter participation for every ethnic minority group (Pacheco et al., 2013). From a trial patient's perspective, the concerns of not being assigned to an effective (or safe) treatment or not achieving a desired outcome could either discourage their participation or result in their withdrawal from a study. Thus, the potential for selection bias is increased and the generalizability of trial conclusions is jeopardized.

Bayesian adaptive designs (BAD) are known for their flexibility in clinical trial design, allowing for modification to the design based on knowledge gained during the study. BADs are also appealing because they alleviate some of the ethical concerns of traditional clinical trial design, namely fixed randomization and sample size. Interestingly, while two-armed designs are the most common in BAD (Lee & Chu, 2012), they have slightly less power than a fixed and balanced design competitor (Coad & Rosenberger, 1999; Berry, 2011; and Connor et al., 2013). In this work, we use a novel optimality assessment to identify the 'best' study design that balances statistical criteria with that of the American Indian community. Since BAD are often less powerful than traditional fixed designs, we also consider a longitudinal variant that uses interim results to adapt the randomization of subjects to treatment (BADL) to improve statistical power. A type of longitudinal design was considered by Cai, Liu, & Yuan (2014); our approach to longitudinal design is similar (see also Kim et al., 2014; and Huang et. al, 2009), but we consider its power improvement and, more importantly, how its design features translate to direct community benefit. This novel approach to evaluation based on both traditional operating characteristics and other subject-focused trial features leads us to an unbalanced two-armed design as the optimal design.

The remaining manuscript is structured as follows: Section 2 describes the challenges that we encounter that motivate this work and how BAD may provide a solution. In Section 3, we describe the trial design, the models of response and accrual, the adaptive randomization procedure, stopping criteria, the simulation procedure, and the method we employ to evaluate multiple criteria simultaneously. Results are presented in Section 4. Discussion and concluding remarks are given in Section 5.

2. MOTIVATION

American Indians (AI) have some of the worst health outcomes of any racial or ethnic group in the U.S. For example, reports of AI smoking rates vary from 32.4% to 40.8%, much higher than African Americans (23.0%–24.3%) and Whites (21.9%–23.6%) (Okuyemi et al, 2004; Redwood et al, 1010; Daley et al, 2010). Similar discrepancies in health care

utilization exist for a wide range of diseases (e.g. breast cancer screening, Espey et al., 2007). Few studies have targeted the treatment of health status and/or disease among AI adults. More controlled studies can fill a major gap by identifying culturally appropriate and effective interventions for heterogeneous AI adult populations.

The rationale of the overall approach to reducing health disparities is reflected by a research funding priority at the National Institutes of Health (NIH, <http://grants.nih.gov/grants/guide/pa-files/PA-14-260.html>). The NIH wants investigators to develop, adapt, and *test* the effectiveness of health promotion and disease prevention interventions in AI populations. An obvious testing strategy is to implement randomized controlled trials in order to perform valid causal inference on the AI population. Because the NIH's goal is to translate research findings to AI communities, the proposed trials must be conducted with AI participants. However, the mistrust of investigators by the AI community has limited the use of randomized trials in the community.

A general solution to this problem is grounded in community-based participatory research (CBPR) (Israel, Schulz, & Parker, 2005). Key concepts that differentiate it from other types of research are the partnership it fosters between the academic institution and the community, the equitable distribution of all aspects of the research process, the shared decision making, and the ownership of data. We argue in this paper that Bayesian Adaptive Designs (BAD) fit under the framework of CBPR and, more importantly, can help to gain the trust of AI communities and improve their overall participation. If an effective treatment is found we can (1) translate this knowledge to the community quickly and (2) improve trust and/or provide clear immediate benefit to the community. Our pursuit of BAD stems in part from discussions with community advisors for our research team and an informal poll we conducted at a community event in which participants chose BAD over a standard randomization in a ratio of 3:1 ($N = 24$).

To illustrate the potential impact of a BAD on the community, consider Figure 1. In planning a CBPR study with a randomized clinical trial to test an experimental intervention, we can plan for N_{max} participants in the study. Using a BAD we can stop the trial early for success using N participants ($< N_{max}$) and use the rest of the resources to offer the better intervention to the $N_{max} - N$ community members. This can potentially put more participants on the better intervention and translate the results faster to the research field and the community.

3. METHODS

Throughout this paper we focus on trials that have a binary endpoint, though similar conclusions apply for continuous, time-to-event, or count data scenarios. In particular, we focus on a trial for smoking cessation. In this trial, we aim to:

1. Examine the acceptability and feasibility of implementing an *individual* telephone-based All Nations Breath of Life (ANBL). This intervention is hypothesized to be better because of its culturally-tailored components.
2. Compare individual telephone ANBL with an individual non-tailored telephone-based program.

3. Disseminate results to the scientific community and the greater AI community.

Our primary efficacy endpoint is the cigarette *quit rate* at 6-months (technically defined as the 30-day point prevalence abstinence from recreational smoking biochemically verified by salivary cotinine level at 6 months post-baseline). Our main hypothesis is that the quit rate in the ANBL program will be significantly better than that in the non-tailored program.

3.1. Design

An overview of the design:

- Model: each arm's 26-week endpoint is binary and is thus modeled with a Bernoulli distribution. Comparisons are made between a design that uses longitudinal information from an interim observation of quit rate at four weeks (BADL), one that uses none of the longitudinal information (BAD), and a fixed design that has a fixed sample size with equal allocation of participants to study arms. The longitudinal model conditions on the four-week quit rate (defined as the 30-day point prevalence abstinence from recreational smoking biochemically verified by salivary cotinine level at four weeks post-baseline), also using a Bernoulli distribution.
- Number of comparative arms: investigation is done using designs with two and three arms.
- Minimum sample size: varies from 250 (two arms) to 330 (three arms);
- Maximum sample size: varies from 500 (two arms) to 660 (three arms);
- Success/stopping criteria: the posterior probability that an arm has the maximum quit rate is greater than some value; the designed cutoff for this probability depends on the number of arms and whether BAD or BADL is used. Variations of the criterion are chosen in order to keep the Type I error rate within a reasonable range (~5 – 10%).
- Interim analysis: After 20 participants are randomized to each arm using fixed, equal randomization, we begin the process of adapting the randomization schedule discussed below. Once 26-week endpoint data on the minimum sample size are collected, we provide a new update of the randomization and assess whether we have achieved the stopping criteria (i.e., identified the best arm). If we have not achieved the stopping criteria, we will continue enrolling the participants until the stopping criteria is achieved or we have reached the maximum sample size.
- Accrual rate: this will vary and will be optimized. The slowest practical accrual rate depends on the maximum sample size; we will provide operating characteristics for accrual rates that are faster and slightly slower than practical in order to investigate a slower effect.

3.2. Model

The primary efficacy endpoint is the number of participants out of n_d participants randomized to treatment arm d who successfully quit which, for modeling purposes, we assume is $Y_{\cdot d} = \sum_{i=1}^{n_d} Y_{id} \sim \text{Bin}(n_d, P_d)$, with $Y_{id} = 1$ if participant i in arm d quits smoking and $Y_{id} = 0$ otherwise. Thus, P_d is the probability of quitting smoking for participants randomized to arm d . We use a log-odds structure for the probability of a quit, where $P_d = e^{\theta_d} / (1 + e^{\theta_d})$. We assume relatively non-informative prior distributions $\theta_d \sim \mathcal{N}(0, 1.82^2)$ for all d . These prior distributions are considered “agnostic” as they are very close to a uniform distribution in the probability space (Beta(1,1) prior distribution).

To evaluate the impact of using an intermediate endpoint for each patient to adapt the randomization of future patients and to make decisions about whether to continue study accrual, a model is proposed that uses the intermediate endpoint longitudinally to update the posterior distribution of a quit. Participants have an interim observation four weeks after randomization at which smoking status is assessed. We label this intermediate endpoint X_{id} for participant i from arm d . The model specifications are common across all arms, including the prior distributions, but separate instances of these models are fit for each arm.

The longitudinal model is a Beta-Binomial distribution such that $P_{d|x} = \Pr(Y_{id} | x) \sim \text{Beta}(\alpha_x, \beta_x)$ where $\alpha_x = \alpha_{1x} + \sum_i \delta_{Y_{id}=1} \delta_{Y_{id}=x_{id}}$ and $\beta_x = \beta_{1x} + \sum_i \delta_{Y_{id}=0} \delta_{Y_{id}=x_{id}}$. δ is a Kronecker delta (Berry & Sanil, 2010), and x is an indicator of the intermediate endpoint ($x = 1$ if quit). The parameters α_x and β_x can be interpreted as the sum of prior expected quits ($\alpha_{1x_{id}}$ where $x_{id} = 1$) and observed interim successes ($x_{id} = 1$) whose final smoking status also indicates quit ($Y_{id} = 1$), and the sum of prior expected failures ($\beta_{1x_{id}}$ where $x_{id} = 0$) and observed interim failures ($x_{id} = 0$) whose final smoking status indicates failure to quit ($Y_{id} = 0$). Weakly informative priors are used: for interim quits ($x = 1$), $\alpha_1 = 3$, $\beta_1 = 1$; for interim failures ($x = 0$), $\alpha_0 = 1$, and $\beta_0 = 3$. This specification represents a belief that, if a participant quits smoking after four weeks, their expected probability of quitting after 26 weeks is 0.75 (based on previous studies). Similarly, a non-quitter at week four has 0.25 probability of quitting at the final endpoint. Both priors have an equivalent sample size of four, making the information weakly informative.

The posterior distribution can be calculated at any interim time. Specifically, we calculate the posterior distribution of the smoking quit rate for the d^{th} treatment, P_d , for Y_d quitters among n_d currently enrolled participants. Markov chain Monte Carlo is used for inference.

In the BAD, we use the posterior probability that arm d is the most effective arm (the largest true smoking quit rate). Calculated for each of D treatment arms, this posterior probability is labeled $\Pr_d(\max) = \Pr(d = \arg \max(P_1, \dots, P_D))$.

3.3. Success Criteria

At the completion of the trial, we determine according to a prespecified threshold of evidence whether we have found the best arm. This decision is made by comparing the probability that an arm is the most effective to γ_S ; we deem the trial a success if, for some arm d , $\Pr_d(\max) > \gamma_S$.

3.4. Response Adaptive Randomization

Response adaptive randomization (RAR) allows us to adjust how we assign participants to treatments during the study based on information gained during the study. We focus the adaptive randomization on finding the best arm, which is labeled as $d_{\max}$. To achieve this goal, we measure the expected additional reduction in variance of adding one more

participant to the d^{th} arm: $I_d = \sqrt{\text{Pr}_d(\max) \text{Var}(\theta_d) / (n_d + 1)}$. I_d is calculated for each of the D arms and then rescaled to a measure of probability that dictates the chances of arm d being assigned to the next participant; specifically, the probability of assigning the next participant to arm d is $r_d = I_d / \sum I_d$. This randomization formula is attractive because the probability of arm d receiving the next participant is proportional to the probability that arm d is the best. However, when this probability is the same across arms, the arm with the higher standard error receives a higher probability in order to gain more information. Note that if r_d is less than 0.05, then we temporarily suspend the d^{th} arm.

3.5. Stopping Criteria

We calculate the probability that an arm is the most effective (stopping criteria) after we have final endpoint data on $N_{\min}$ participants, and this calculation is repeated every eight weeks. We stop the trial early if there is an arm such that $\text{Pr}(d = d_{\max}) > \gamma_E$, where γ_E depends on the design (e.g., number of arms).

3.6. Various Accrual Patterns

The accrual rate for clinical trials is a critical component of a design, and should be treated as a design component. In fact, it is well known that investigators in medical research tend to over-promise how fast they can recruit and underestimate how long it takes. It is estimated that more than 80% of clinical trials run past the original estimated timeframe in order to meet accrual goals (e.g. van der Wouden et al., 2007). Investigators' underestimation of the sample variability in the accrual timeframe is a potential reason for this problem, even in cases for which an investigator is very good at estimating the *expected* accrual timeframe.

However, as pointed out by Gajewski et al. (in press), the smallest sample size for a BAD is the one for which the slowest accrual rate is feasible. This is attributed to the fact that use of the adaptive engine is maximized to drive knowledge about which treatment is best. Thus, an optimal design balances the risk of finishing the trial late with the information gained from slowing the accrual to the lowest possible rate. We propose to accomplish this by designing our risk so that we are 95% sure we finish the trial on time when there are no differences between the treatment arms, a scenario which requires the largest sample size, $N_{\max}$. Designing the BAD in this way allows the scenarios for which treatment arm differences exist: (1) to finish faster than planned; (2) to place more participants on the better-performing arm; and (3) to use less of the planned participants for the investigational piece of the study. This allows us to translate our findings to the literature faster and use the trial's remaining resources to implement the best treatment within the community. We let N_{com} be the total number of participants placed on the best treatment. When the trial stops early for efficacy (e.g., $N < N_{\max}$), N_{com} includes the investigational portion of the implementation and the post-trial work once we have identified the best treatment. Thus, in our aim to

optimize N_{com} we assume that any participants not included in the investigational portion of the trial (i.e., $N_{max} - N$) will receive the treatment identified as best.

Needing an accrual rate goal, we use the accrual model in Gajewski, Simon, and Carlson (2008) and Jiang, Simon, Mayo, and Gajewski (2015). Define the average time elapsed between subsequent participants accrued to trial as $1/\lambda$ (in weeks). A researcher wants to be fairly certain (e.g., 95% certain) to accrue N participants in T weeks. Placing a gamma distribution on T , $T \sim \Gamma(N, 1/\lambda)$, we can solve for $1/\lambda$ in $0.95 = \int_0^T \lambda^N / (\Gamma(N)) x^{N-1} e^{-x\lambda} dx$. The solution $1/\lambda$ allows us to be 95% certain to finish the trial before T weeks (λ participants/week). Thus, λ is the lowest average accrual rate, and we investigate how increasing this impacts usual BAD optimization parameters (extending the work of Gajewski et al., in press) including trial size, duration, allocation to arms, and number of community members receiving the best arm.

3.7. Virtual Responses (i.e., Effect Sizes for Null and Alternative Hypotheses)

All of these optimization parameters vary depending on the factor combination. The primary endpoint is the rate of smoking cessation (quit) at 26 weeks, which is assumed to be 15% for the standard program. The experimental program is assumed to have a quit rate of 20%, 25%, or 35%. The longitudinal data patterns are determined based on the percentage of participants who quit smoking. Using the notation from above, we assume $\Pr(Y_{id} = 1 | x = 0) = 0.12$, $\Pr(Y_{id} = 1 | x = 1) = 0.8$, and $\Pr(X = 1) = 0.25$. In other words, we assume that those who quit at four weeks have an 80% chance of staying quit by 26 weeks and those who don't have a 12% chance of quitting at 26 weeks. The last says that 25% of the participants quit at four weeks.

3.8. Factors for Simulation investigating a Binary Endpoint

To simulate across different clinical trial designs, we use a three-way factorial simulation design that is motivated from the tANBL survey (See Table 1). The first factor varies number of arms (2 or 3) and the second varies the use of longitudinal data (yes/no). The longitudinal data are intended to move the knowledge of the endpoint a little closer to optimal randomization which should increase the optimization of the BAD design parameters. The minimum sample size, frequency of interim analysis, and stopping rules for success are dictated by having acceptable Type I error rates ($< 10\%$). The maximum sample size is dictated by frequentist power (70 – 100%) for a fixed design. FACTS™ software generates random numbers for these simulations and calculates the operating characteristics of the trial. The basic simulation is 1,000 iterations for each combination of the five factors. We compare the results to a fixed design (fixed sample size and equal allocation).

3.9. Investigational Outcomes of the Simulations

Balanced two-arm designs are well known for being more powerful than unbalanced designs (fixed). From that vantage point, a fixed design has better properties than a BAD when comparing two treatments. However, other design characteristics, such as the number of community members receiving the better treatment and the overall number of participants that quit smoking, are also important for evaluating competing designs. For this reason, we

view this as a multiple “output” problem. Data Envelopment Analysis (DEA) is a nonparametric approach to estimating efficiency that uses mathematical programming to estimate the relative performance of business decision making units (DMU) in the presence of multiple inputs and outputs (Cooper, Seiford, & Tone, 2007). Simply put, a DMU is an entity that can be evaluated for efficiency by measuring its production (outputs) relative to its resources used (inputs). DEA estimates a “frontier” on which the most efficient DMUs reside. In our case, the type of design is the DMU and we use DEA to measure the efficiency of the competing designs (see Mandelblatt et al., 2009). The primary benefit of DEA in this context is its lack of assumptions regarding the relationships between inputs and outputs which permits complex relationships between design characteristics measured in different units. The FEAR package (Wilson, 2006) in R program is used to estimate the frontier and calculate efficiencies for each design. This allows us to identify the most efficient design (i.e., fixed, BAD, or BAD with longitudinal) while incorporating a compromise of power, size, community contribution, and quit rates of the different designs into the comparison.

Overall expected sample size, expected allocation of the sample size across arms, power to detect the best arm, and average trial duration are the key optimization parameters. We evaluate these across scenarios and accrual rates. Additionally, we investigate the total number of participants who are placed on the best treatment (N_{com}) and total number of participants who quit smoking.

4. RESULTS

We vary N_{max} , the maximum sample size for the trials, to be 500 for the two arm studies but increase it to 660 for the study with three arms. These values are used to determine the minimum accrual rates that give us enough time to enroll and collect final endpoint data within a four-year time period. For 500 and 660 participants, the minimum accrual rates are determined to be $\lambda = 3.1$ and 3.8 participants per week, respectively.

4.1 Two arms

In Table 2, we have reported the operating characteristics for the BAD design for various null and alternative hypotheses to determine the power, sample size, participant allocation, time (duration), number of community members receiving the better treatment, and the total number of participants who will quit smoking for our study. We created several scenarios for quit rates. We performed 16 sets of trial simulations based on the various combinations of response and accrual rates shown in Table 2. Each set involved 1,000 trial simulations. We highlight two scenarios, both at $\lambda = 3.1$ * participants accrued per week, the slowest acceptable rate. The first uses what we believe is the most likely response ($\theta_1 = 0.25$ and $\theta_2 = 0.15$). If there is a best group in terms of quit rate, we estimated (identified) that 64% of the simulated trials had early success and 7% had late success. This trial scenario had 71% power, an average sample size of 394 (65% of these in the winning group), with 363 community members getting the better intervention and 111 community members who quit smoking. The average duration of this trial scenario was 155 weeks. While a conventional equal randomization trial (fixed) would have higher power (85%), it would be larger (500 participants), slower (162 weeks), have a lower rate of participants on the winning group

(50%), and only 100 community members would be expected to quit smoking. The second highly unlikely scenario serves as our null hypothesis ($\theta_1 = 0.15$ and $\theta_2 = 0.15$ in Table 2). In this scenario, there are no differences in quit rates among the groups. Therefore, the extent to which this scenario is “successful” actually reflects our Type I error rate. For this scenario, we estimated (identified) that 7% of the simulated trials had early success and 0.5% had late success. Thus, this trial scenario produced an appropriate expected Type I error ($\alpha = 7.5\%$). The sample size of this scenario was, on average, 489 participants (equally allocated across groups). The average duration of the trials under this scenario was 188 weeks. While the BAD design has acceptable Type I error rates and the potential impact to significantly expedite the findings to the research community and clinical practice, its power can be improved enough to be competitive with the fixed design.

Therefore, the improvement was accomplished by using more information by applying the longitudinal design (displayed in Table 3). Again, consider the most likely response ($\theta_1 = 0.25$ and $\theta_2 = 0.15$) at $\lambda = 3.1$ participants accrued per week. If there is a best group in terms of quit rate, we estimated (identified) that 73% of the simulated trials had early success and 8% had late success. This trial scenario had 80% power, an average sample size of 376 (63% of these in the winning group), and 111 community members who quit smoking. The average duration of this trial scenario was 149 weeks. Therefore, incorporating the longitudinal design gives us power much closer to a conventional design but provides decisions using fewer participants and even faster than the BAD.

In both BAD and BADL we can see that slower accrual gives better trial outcomes in terms of participants used, community members on better intervention, and total participants who quit. However, it is impractical to go slower given the timeframe of funding from the sponsor. Conversely, going faster gives poorer trial results in terms of impact on the community.

What, then, is the best design? First, consider a graphical analysis (Figures 2 and 3). In Figure 2, we show the Type I error rate and power for the fixed, BAD, and BADL designs at the minimum accrual rate of $\lambda = 3.1$ participants per week. Power is for the average effect sizes from the three alternative hypotheses. We can see that the three have similar Type I errors and the BAD has smaller power than the comparable fixed and BADL. However, a clear separation is shown for the other operating characteristics (Figure 3). We can see that both BAD have better results than the fixed design. Relative to the fixed design, the BAD put a higher percentage of participants on the better intervention, are shorter in duration (by approximately one year!), place more community members on the better intervention (by 100!), and more community members quit smoking (close to 20!). These are much better overall results favoring the BAD over the fixed design. Second, using a strategy that balances benefits and harm (Mandelblatt et al., 2009), we use an efficiency analysis. We identify the design(s) that have the best tradeoff of power and community impact as measured by the number of participants receiving the better intervention and the number that will quit smoking. As we shall see later, confirming the literature, the studies with two arms are the only cases where the power of the fixed design is higher than BAD. The top of Figure 4 displays the trade-off of input ($1 - \text{power}$) and outputs (expected quit and expected community members on the better treatment) as a function of several designs with two arms

for an alternative hypothesis that indicates a small effect ($\theta_1 = 0.25$ and $\theta_2 = 0.15$). The most efficient designs are those that have the highest ratio of outputs to input, thereby placing them at the top right of the graph. Designs that are most efficient define the ‘best practices frontier,’ which is represented on this graph by lines connecting the most efficient design (BAD2.5L) to the y - and x -axes. The most efficient design among those included in Figure 4 is the longitudinal BAD with $\lambda = 2.5$, an accrual rate that is too slow to be practical. The second most efficient design is the slowest practical longitudinal design (BADL with $\lambda = 3.1$). The third slowest longitudinal BAD (with $\lambda = 5$) is the third most efficient design. Ranked behind all of these are the faster accruing designs, non-longitudinal designs, and those that are fixed. The bottom graph of Figure 4 is the efficiency score at each design averaged across small, moderate, and large effects. An efficiency score close to 1 indicates a highly efficient design—those designs that lie on the frontier have a score of 1. All others will have fractional scores that represent their efficiency relative to the frontier. Again, the longitudinal BAD design with slowest accrual is most efficient. The second and third most efficient are longitudinal designs with faster accrual rates. The fixed design, despite power advantages, is the fourth best.

4.2. Three arms

In Table 4, we have reported operating characteristics for the three-arm case and confirm clear BAD superiority over the fixed design. We also highlight two scenarios, both at the $\lambda = 3.8^*$ participants per week accrual rate (the slowest acceptable). The first scenario is the most likely response ($\theta_1 = 0.25$, $\theta_2 = 0.15$, and $\theta_3 = 0.15$). If there is a best group in terms of quit rate, we estimated (identified) that 77% of the simulated trials had early success and 6% had late success. This trial scenario had 83% power, an average sample size of 484 (49% of these in the winning group), 414 community members on the better intervention, and 140 community members who quit smoking. The average duration of this trial scenario was 154 weeks. A conventional equal randomization trial would have a lower power (64%), require more participants (660 participants), take longer to make a decision (200 weeks), have a lower percentage of participants on the winning arm (33%), and result in only 121 community members who quit smoking. The BAD design has the potential impact to significantly expedite the dissemination of findings to the research community, communities who may wish to implement the program, and clinical practice and its power is superior to the fixed design.

5. DISCUSSION

Over the past year we have developed many more BAD designs for our CBPR studies. We have submitted proposals for investigating the improvement in culturally-tailored smoking cessation and weight-loss programs, among others. We have also elicited the opinion of a small portion of the community of AI regarding fixed versus adaptive designs. We recently found that a BAD was preferred over a fixed design in a 3:1 ratio by AI community members, from 24 AI surveyed at a large community event. The presentation of the comparison did not include the idea of transferring the rest of the resources to the community, but we believe an even larger discrepancy occurs if we make such a presentation. Specifically, we can now say using the slowest feasible accrual rate, that for a

500-person study we can, on average, place many of these participants on the better smoking program without big sacrifices in power, helping more community members quit smoking immediately.

We have been reminded that the most common BAD (two arms) do not provide better power than fixed randomization (Berry, 2011; Connor et al., 2013) and our simulations are consistent, but this gap closes when using a longitudinal design. We are also consistent with the fact that with more than two arms BAD sees gains in power over the fixed design (Connor et al., 2013). However, when attempting to gain trust from a community we argue that power should not be the only parameter of interest. Translation of information that directly affects a community should also be a consideration, and from this regard we clearly demonstrate BAD is a very good choice.

By designing BAD that use the slowest practical accrual patterns and incorporate a longitudinal data model, we can have optimal or near optimal design parameters for experimental sample size, distribution of sample size to best treatment, very close to optimal power (or better), and translate the most resources to the community more quickly. We hope that by designing more BAD in CBPR involving AI, researchers gain more of the communities' trust by providing immediate benefit from the research with transfer of better health opportunities to the community. In fact, we can now say in the consent form that "as a community, overall, more participants will most likely be placed on the better arm." Such a statement could improve enrollment rates and trust from the community. BAD may also provide more appealing choices to other communities with trust deficits. At a minimum, when conducting CBPR, BAD should be presented to communities as options to determine community preference. We have demonstrated this advantage for the binary case (including two and three arms) and our next step is to demonstrate this for continuous and survival endpoints.

The results of our study here can easily generalize to other applications. Other examples may also have different follow-up times, may incorporate stopping rules for futility or models for handling drop-outs, and/or choose other trial characteristics that are important to optimize, for example cost.

Acknowledgments

Partial funding for all the authors, except the second, comes from a grant from the USA NIH, National Institute on Minority Health and Health Disparities (5P20MD004805).

REFERENCES

1. Berry DA. Adaptive clinical trials: The promise and the caution. *J Clin Oncol*. 2011; 29(6):606–609. [PubMed: 21172875]
2. Berry, S., Sanil, A. FACTS™ Dose finding: single endpoint engine specification. Newton, MA: Tessela; 2010.
3. Berry, SM., Carlin, BP., Lee, JJ., Muller, P. Bayesian Adaptive Methods for Clinical Trials. New York: CRC Press; 2011.
4. Berry SM, Spinelli W, Littman GS, Liang JZ, Fardipour P, Berry DA, Lewis RL, Krams M. A Bayesian dose-finding trial with adaptive dose expansion to flexibly assess efficacy and safety of an investigational drug. *Clinical Trials*. 2010; 7:121–121. [PubMed: 20338905]

5. Cai C, Liuc S, Yuanc Y. A Bayesian design for phase II clinical trials with delayed responses based on multiple imputation. *Statistics in Medicine*. (in press) early view.
6. Coad D, Rosenberger W. A comparison of the randomized play-the-winner rule and the triangular test for clinical trials with binary responses. *Statistics in Medicine*. 1999; 18:761–769. [PubMed: 10327525]
7. Connor JT, Luce BR, Broglio KR, Ishak KJ, Mullins CD, Vanness DJ, Fleurence R, Saunders E, Davis BR. RE-ADAPT study Do Bayesian adaptive trials offer advantages for comparative effectiveness research? Protocol for the RE-ADAPT study. *Clinical Trials*. 2013; 10:807–827. [PubMed: 23983160]
8. Cooper, W., Seiford, L., Tone, K. *Data envelopment analysis: a comprehensive text with modes, applications, references and DEA-solver software*. London, UK: Kluwer Academic Publishers; 2007.
9. Daley CM, Greiner KA, Nazir N, Daley SM, Solomon CL, Braiuca SL, Smith TE, Choi WS. All Nations Breath of Life: using community-based participatory research to address health disparities in cigarette smoking among American Indians. *Ethn Dis*. 2010; 20(4):334–338. [PubMed: 21305818]
10. Espey DK, Wu XC, Swan J, Wiggins C, Jim MA, Ward E, et al. Annual report to the nation on the status of cancer, 1975–2004, featuring cancer in American Indians and Alaska Natives. *Cancer*. 2007; 110(10):2119–2152. [PubMed: 17939129]
11. Gajewski BJ, Berry SM, Pasnoor M, Dimachkie M, Herbelin L, Barohn R. Building Efficient Comparative Effectiveness Trials through Adaptive Designs, Utility Functions, and Accrual Rate Optimization: Finding the Sweet Spot. *Statistics in Medicine*. (in press).
12. Gajewski B, Simon S, Carlson S. Predicting Accrual in Clinical Trials with Bayesian Posterior Predictive Distributions. *Statistics in Medicine*. 2008; 27(13):2328–2340. [PubMed: 17979152]
13. Huang XL, Ning J, Li YS, Estey E, Issa JP, Berry DA. Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Statistics in Medicine*. 2009; 28:1680–1689. [PubMed: 19326367]
14. Inoue LYT, Thall PF, Berry DA. Seamlessly expanding a randomized phase II trial to phase III. *Biometrics*. 2002; 58:823–831. [PubMed: 12495136]
15. Israel, EE., Schulz, AJ., Parker, EA. *Methods in Community-Based Participatory Research for Health*. San Francisco: Jossey-Bass; 2005.
16. Jennison, C., Turnbull, BW. *Group Sequential Methods with Applications to Clinical Trials*. New York: Chapman & Hall/CRS; 2000.
17. Jiang, Simon, Mayo, Gajewski. Performance of Constant Accrual Model and Alternatives on Clinical Data and Simulation. *Statistics in Medicine*. 2015; 34(4):613–629. [PubMed: 25376910]
18. Kelly PJ, Stallard N, Todd S. An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics*. 2005; 15:641–658. [PubMed: 16022169]
19. Kim MO, Liu C, Hu F, Lee JJ. Outcome-adaptive randomization for a delayed outcome with a short-term predictor: imputation-based designs. *Statistics in Medicine*. 2014; 33(23):4029–4042. [PubMed: 24889540]
20. Lee JJ, Chu CT. Bayesian clinical trials in action. *Statistics in Medicine*. 2012; 31(25):2955–2972. [PubMed: 22711340]
21. Mandelblatt JS, Cronin KA, Bailey S, Berry DA, Koning H, Draisma G, Huang H, Lee SJ, Munsell M, Plevritis SK, Ravdin P, Schechter CB, Sigal B, Stoto MA, Stout NK, van Ravesteyn NT, Venier J, Zelen M, Feuer EJ. for the Breast Cancer Working Group of the Cancer Intervention and Surveillance Modeling Network (CISNET). Effects of Mammography Screening Under Different Screening Schedules: Model Estimates of Potential Benefits and Harms. *Ann Intern Med*. 2009; 151:738–747. [PubMed: 19920274]
22. Okuyemi, K., Cox, LS., Choi, WS., Ahluwalia, JS. *Smoking cessation in US Wthnic Minority Populations*. Vanguard Conference: Building Success in Smoking Cessation; San Francisco. 2004.
23. Pacheco CM, Daley SM, Brown T, Filippi M, Greiner KA, Daley CM. Moving Forward: Breaking the cycle of mistrust between American Indians and researchers. *American Journal of Public Health*. 2013; 103(12):2152–2159. [PubMed: 24134368]

24. Redwood D, Lanier AP, Renner C. Differences in cigarette and smokeless tobacco use among American Indian and Alaska Native people in Alaska and the Southwest United States. *Nicotine and Tobacco Research*. 2010; 12(7):791–796. [PubMed: 20525781]
25. National Medical Association. Recommendations of the clinical trials consensus panel - National Medical Association. *Journal of the National Medical Association*. 2000; 92(10):464–471. [PubMed: 11105726]
26. van der Wouden JC, Blankenstein AH, Huibers MJ, van der Windt DA, Stalman WA, Verhagen AP. Survey among 78 studies showed that Lasagna's law holds in Dutch primary care research. *J Clin Epidemiol*. 2007; 60(8):819–824. [PubMed: 17606178]
27. Wilson PW. FEAR: A Software Package for Frontier Efficiency Analysis with R. 2006 Retrieved from <http://www.clemson.edu/economics/faculty/wilson/software/fear/Paper/fear.pdf>.

Figure 1.

We have planned a CBPR study of N_{max} participants using a randomized clinical trial to test an experimental intervention (e.g., smoking cessation). Using a BAD, we can stop the trial early for success using $N - N_{max}$ participants and use the rest of the resources to offer the better intervention to the $N_{max} - N$ community members. This can potentially put more participants on the better intervention and translate the results faster to the research field and the community. In fact, if N_1 is the number of participants in the better intervention, this design places the number of community members on the better arm at $N_{com} = N_1 + (N_{max} - N)$. Pictures come from: (1) http://obssr.od.nih.gov/training_and_education/annual_Randomized_Clinical_Trials_course/RCT_info.aspx and (2) <http://newsinhealth.nih.gov/issue/apr2014/feature1>.

Figure 2.

Two arms. Type I error rate (top) and power (bottom) for the fixed, Bayesian adaptive (BAD), and Bayesian adaptive with longitudinal (BADL) designs. Power is for the average effect sizes from the three alternative hypotheses. The accrual rate of $\lambda = 3.1^*$ participants per week is the slowest acceptable rate.

Figure 3.

Two arms. Operating characteristics for fixed, Bayesian adaptive (BAD), and Bayesian adaptive with longitudinal (BADL) designs. $E(\% \text{ Best})$ is the average percent allocation of trial participants to the better intervention, with average overall trial size $E(N)$. $E(T)$ is the average duration of the trial in weeks. $E(N_{\text{com}})$ is the average total number of community members receiving the new and better intervention (out of 500). $E(\text{Quit})$ is the expected number of participants that quit smoking for the entire study.

Figure 4.

Two arms efficiency analysis. Fixed, Bayesian adaptive (BAD), and Bayesian adaptive with longitudinal (BADL) designs are shown, each having a number that designates the accrual rate. The top figure displays the trade-off of input $(1 - Power)$ and outputs (the number of participants expected to quit, $E(Quit)$, and the expected number of community members receiving the better intervention, $E(N_{com})$) as a function of several designs with two arms at small effect. The solid line represents the fully-efficient frontier. The bottom figure shows the efficiency for each design averaged across small, moderate, and large effects. The

longitudinal BAD design with slowest accrual is most efficient. The second and third most efficient are longitudinal designs with faster accrual rates. The fixed design, despite power advantages, is the fourth best.

Table 1

Here are the various designs for investigating tANBL as an effective smoking cessation program.

Arms	Longitudinal	γ_E	$N_{\min}$	$N_{\max}$	Slowest Allowable Accrual Rate
2	No	0.990	250	500	3.1
2	Yes	0.980	250	500	3.1
2	Yes	0.975	250	500	3.1
3	No	0.950	330	660	3.8
3	No	0.950	330	660	3.8

The stopping for success probability $\Pr(d = d_{\max}) > \gamma_E$ varies by design. The stopping and evaluation rules were chosen to keep the Type I error rates at around 5–10%.

Two-arm BAD scenario. The fixed design's power at a Type I error of 0.07 (two-sided) is 0.37, 0.84, and 1.0 for the three alternative hypothesis scenarios. θ_1 is the true smoking rate of the experimental intervention and the control rate is $\theta_2 = 0.15$.

Table 2

θ_1	Rate λ	Early $P(S)$	Late $P(S)$	Power	$E(N_1)$	$E(N_2)$	$E(N)$	$E(T)$	$E(N_{com})$	$E(Quit)$
.15	2.5	0.071	0.007	0.078	243	247	489	225	254	75
	3.1*	0.070	0.005	0.075	244	245	489	188	255	75
	5	0.082	0.010	0.092	244	244	487	127	257	75
	10	0.053	0.020	0.073	247	246	493	79	254	75
.20	2.5	0.247	0.035	0.282	275	184	459	213	316	91
	3.1*	0.231	0.035	0.286	276	188	463	179	312	91
	5	0.211	0.056	0.267	274	195	469	123	305	90
	10	0.147	0.128	0.275	263	221	484	78	279	89
.25	2.5	0.664	0.074	0.738	254	130	384	181	370	112
	3.1*	0.637	0.072	0.709	257	137	394	155	363	111
	5	0.562	0.136	0.698	258	154	412	110	346	110
	10	0.331	0.392	0.723	263	202	465	75	298	105
.35	2.5	0.994	0.005	0.999	203	72	276	136	427	160
	3.1*	0.992	0.006	0.998	202	79	282	117	421	159
	5	0.984	0.013	0.997	196	104	301	86	395	154
	10	0.833	0.166	0.999	230	173	402	67	328	141

The accrual rate of $\lambda = 3.1$ * participants accrued per week is the slowest acceptable rate. $H(S)$ is the probability of trial success; $H(N_1)$ and $H(N_2)$ are the average number of participants allocated to the experimental and control arms, respectively, with average total sample size $E(N)$. $E(T)$ is the average duration of the trial in weeks. $E(N_{com})$ is the average total number of community members receiving the new and better intervention (out of 500), and $E(Quit)$ is the expected number of participants that quit smoking for the entire study (out of 500). For the fixed design, $H(N_{com}) = 250$, and $H(Quit) = 88$, 100, and 125 for the three alternative hypothesis scenarios.

Table 3

Two-arm BAD longitudinal scenario. The fixed design's power at a Type I error of 0.07 (two-sided) is 0.37, 0.84, and 1.0 for the three alternative hypothesis scenarios.

θ_1	Rate λ	Early $P(S)$	Late $P(S)$	Power	$E(N_1)$	$E(N_2)$	$E(N)$	$E(T)$	$E(N_{com})$	$E(Quit)$
.15	2.5	0.078	0.015	0.093	245	243	488	225	257	75
	3.1*	0.070	0.020	0.090	244	245	489	188	255	75
	5	0.057	0.026	0.083	247	246	493	129	254	75
	10	0.005	0.052	0.057	250	250	500	79	250	75
.20	2.5	0.328	0.052	0.380	160	186	447	208	314	91
	3.1*	0.304	0.068	0.372	264	190	453	175	310	91
	5	0.229	0.129	0.358	272	200	471	123	300	90
	10	0.038	0.254	0.292	272	226	498	79	274	89
.25	2.5	0.771	0.059	0.830	231	129	360	171	371	112
	3.1*	0.728	0.076	0.804	238	138	376	149	362	111
	5	0.661	0.153	0.814	248	155	403	108	345	110
	10	0.247	0.546	0.793	280	203	483	77	297	105
.35	2.5	0.999	0.001	1.000	194	74	268	133	426	160
	3.1*	0.999	0.001	1.000	192	80	272	114	420	159
	5	0.991	0.009	1.000	187	104	291	84	396	154
	10	0.892	0.107	0.999	238	165	403	67	335	142

The accrual rate of $\lambda = 3.1$ * participants accrued per week is the slowest acceptable rate. θ_1 is the true smoking rate of the experimental intervention and the control rate is $\theta_2 = 0.15$. $R(S)$ is the probability of success; $E(N_1)$ and $E(N_2)$ are the average number of participants allocated to the experimental and control arms, respectively, with average total sample size $E(N)$. $E(T)$ is the average duration of the trial in weeks. $E(N_{com})$ is the average total number of community members receiving the new and better intervention (out of 500), and $E(Quit)$ is the expected number of participants that quit smoking for the entire study (out of 500). For fixed design $E(N_{com}) = 250$, and $E(Quit) = 88$, 100, and 125 for the three alternative hypothesis scenarios.

Table 4

Three arms BAD scenario. The fixed design's power at a Type I error of 0.07–Bonferroni adjusted—(two-sided) is 0.19, 0.64, and 1.0 for the three alternative hypothesis scenarios.

θ_1	Rate λ	Early $P(S)$	Late $P(S)$	$P(S_2)$	$E(N_1)$	$E(N_2)$	$E(N_3)$	$E(N)$	$E(T)$	$E(N_{com})$	$E(Quit)$
.15	2.5	0.080	0.003	0.031	215	213	216	644	287	229	99
	3.8*	0.052	0.008	0.020	219	215	214	649	200	227	99
	5	0.050	0.007	0.024	218	216	216	651	160	226	99
.20	10	0.052	0.013	0.026	216	218	215	649	95	229	99
	2.5	0.282	0.036	0.310	170	260	171	595	270	319	115
	3.8*	0.256	0.063	0.301	175	261	174	609	189	311	115
.25	5	0.254	0.058	0.303	176	259	176	611	151	308	114
	10	0.176	0.117	0.284	191	250	192	633	93	277	113
	2.5	0.807	0.038	0.845	117	234	114	469	213	429	142
.35	3.8*	0.772	0.057	0.823	123	237	123	484	154	414	140
	5	0.728	0.105	0.829	128	240	129	498	127	403	139
	10	0.484	0.294	0.776	167	247	165	579	86	328	132
.45	2.5	1.000	0.000	1.000	66	215	66	346	165	528	205
	3.8*	0.999	0.001	1.000	72	209	74	356	120	514	202
	5	1.000	0.000	1.000	84	201	84	369	100	492	197
.55	10	0.967	0.033	1.000	128	195	128	451	71	404	180

The accrual rate of $\lambda = 3.8^*$ participants per week is the slowest acceptable rate. θ_1 is the true smoking rate of the experimental intervention, a second intervention has a quit rate of $\theta_2 = 0.15$, and the control rate is $\theta_3 = 0.15$. $P(S)$ is the probability of success; $P(S_2)$ is the probability the best arm is successful; $E(N_1)$, $E(N_2)$, and $E(N_3)$ are the average number of participants allocated to the experimental, second experimental, and control arms, respectively, with average total sample size $E(N)$. $E(T)$ is the average duration of the trial in weeks. $E(N_{com})$ is the average total number of community members receiving the new and better intervention (out of 660), and $E(Quit)$ is the expected number of participants that quit smoking for the entire study (out of 660). For fixed design $E(N_{com}) = 220$, and $E(Quit) = 110$, 121, and 143 for the three alternative hypothesis scenarios.