Two-stage Adaptive Randomization for Delayed Response in Clinical Trials

Guosheng Yin

Department of Statistics and Actuarial Science
The University of Hong Kong

Joint work with J. Xu

PSI and RSS Journal Club
May 21, 2015

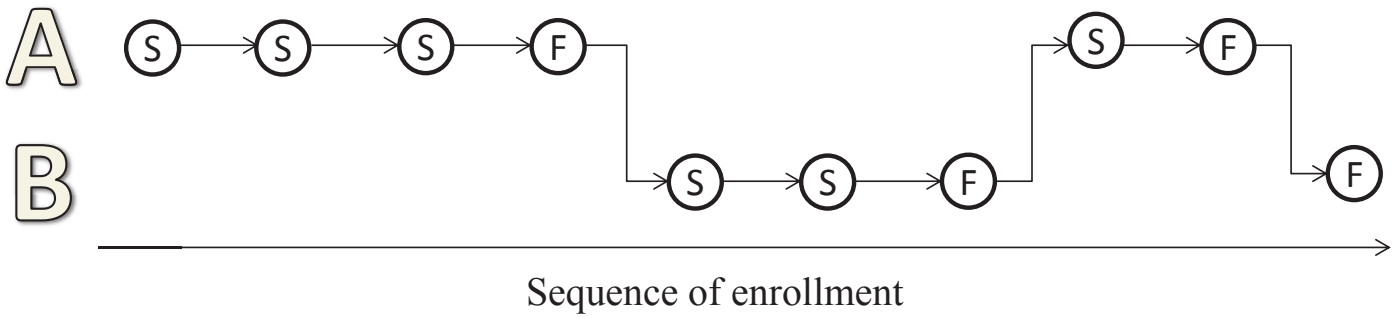
Equal/Adaptive Randomization

- In a clinical trial with multiple treatments, the goal is to identify the superior treatment quickly, as well as treating patients effectively.
- Equal randomization (ER) is a simple and efficient way for patient allocation.
- Response-based adaptive randomization (AR) tends to assign more patients to better treatments based on the information accumulated in the trial.
- Prior to the implementation of AR, a prerun of ER is typically used to stabilize the parameter estimates.
- However, it is not clear how large the prerun sample size should be, and it is often chosen arbitrarily in practice.

Adaptive Randomization

- Pioneering work can be traced back to Thompson (1933), Robbins (1952), and Feldman (1962) etc.
- Play-the-winner rule (Zelen, 1969): Continue using the same treatment if a success response is observed; otherwise switch to the other treatment.
- Randomized play-the-winner rule (Wei and Durham, 1978): A higher randomization probability is given to the treatment that has produced a success response.
- Bandit problems and Bayesian adaptive randomization (Berry and Eick, 1995).

Play-the-Winner Rule



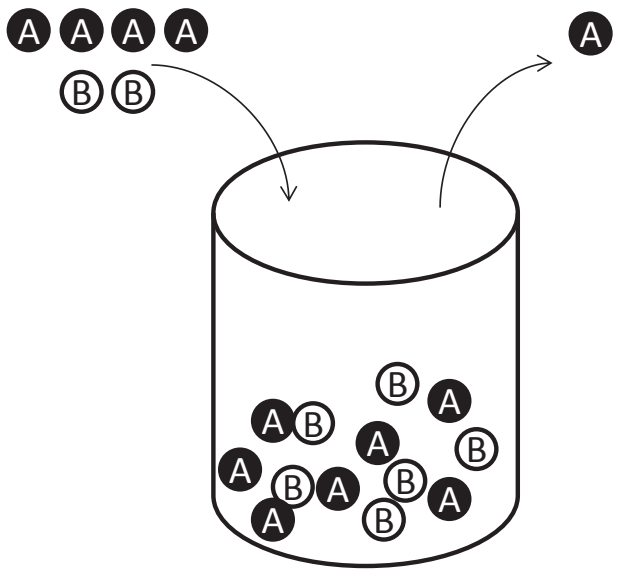


Figure 1: Play-the-winner rule and urn model with treatments A and B.

Optimal Allocation in AR (Binary Data)

- We can calculate the optimal allocation ratio by minimizing the variance (equivalently, maximizing power), or by minimizing the expected number of nonresponders in a trial.
- Let p_1 and p_2 denote the response rates of treatments 1 and 2.
- By minimizing the variance of the difference between $\hat{p}_1$ and $\hat{p}_2$, the allocation ratio between arm 1 and arm 2 is

$$\frac{\sqrt{p_1(1-p_1)}}{\sqrt{p_2(1-p_2)}},$$

  which is known as Neyman's allocation.
- By minimizing the number of nonresponders while fixing the variance (Rosenberger et al., 2001), the allocation ratio is

$$\frac{\sqrt{p_1}}{\sqrt{p_2}}.$$

Continuous Data

- For continuous data, let μ_1 and μ_2 denote the means of two normal distributions, and let σ_1^2 and σ_2^2 denote the corresponding variances.
- Neyman's allocation ratio is

$$\frac{\sigma_1}{\sigma_2},$$

  which minimizes the variance.
- For the case where a smaller response is preferred, Zhang and Rosenberger (2005) proposed an optimal allocation ratio of

$$\frac{\sigma_1\sqrt{\mu_2}}{\sigma_2\sqrt{\mu_1}},$$

  by minimizing the total expected response from all patients.

Bayesian Adaptive Randomization

- **In the Bayesian approach, we may naturally assign patients to treatment 1 with a probability of**

$$\lambda = \Pr(p_1 > p_2 | y_1, y_2),$$

  where y_1 and y_2 represent the accumulated data in the two arms (Yin, 2012).
- **By comparing the posterior distributions of p_1 and p_2, it automatically accounts for both the point and variance estimates of the treatment response rates.**

Bayesian Estimates (Early vs. Late Stages)

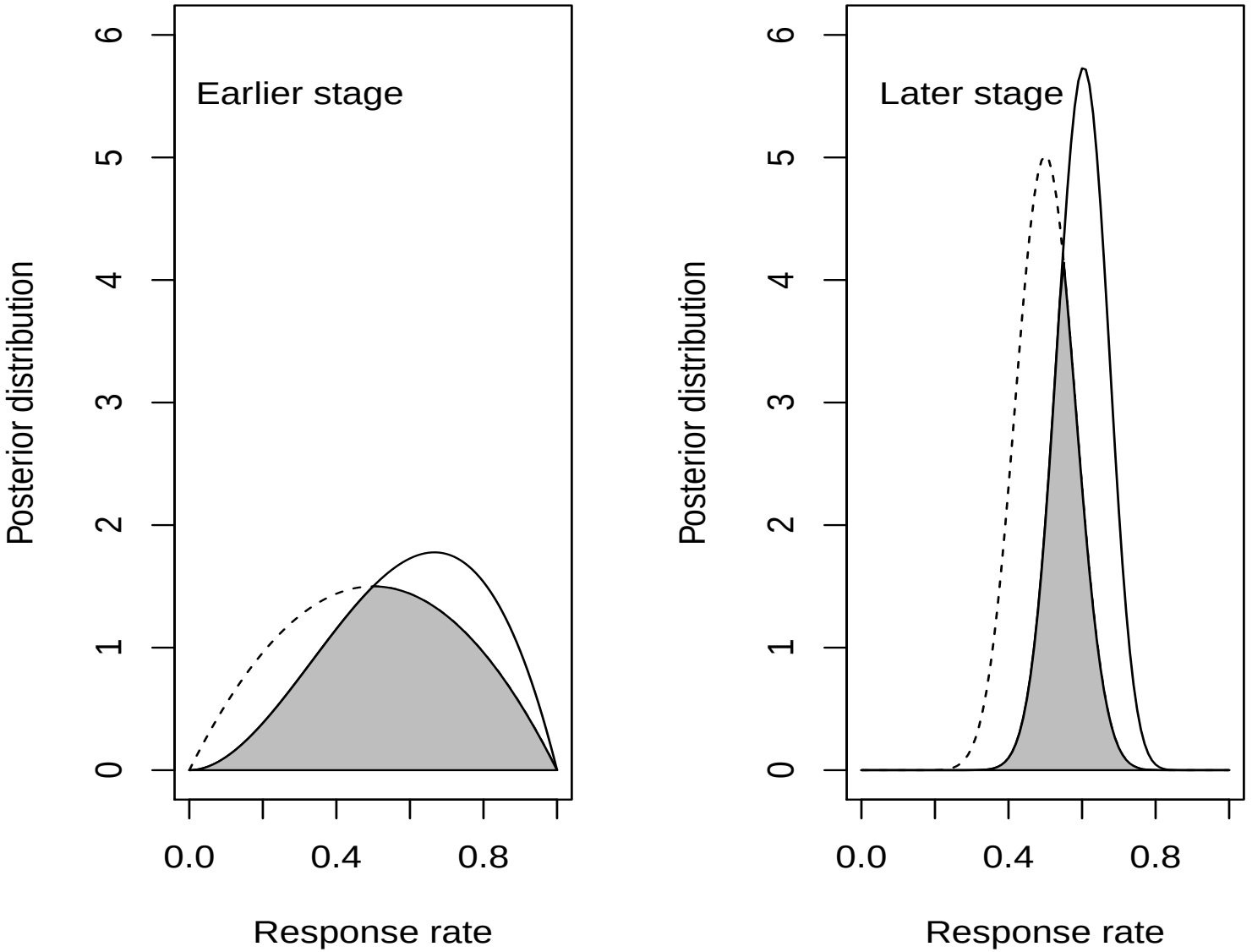


Figure 2: Posterior distributions of the response rates at the earlier and later stages of a trial.

Power Transformation

- **We can explore a class of randomization probabilities,**

$$\pi(\lambda, \gamma) = \frac{\lambda^\gamma}{\lambda^\gamma + (1-\lambda)^\gamma}.$$

- **If $\gamma = 0$, the randomization scheme reduces to ER with an equal assigning probability of 0.5 regardless of the value of λ; and if $\gamma = 1$, $\pi(\lambda, \gamma) = \lambda$.**
- **It may depend on the accumulating sample size n,**

$$\gamma_n = \frac{n}{2N},$$

where N is the total sample size.

Delayed Response with $\tau = 6a$

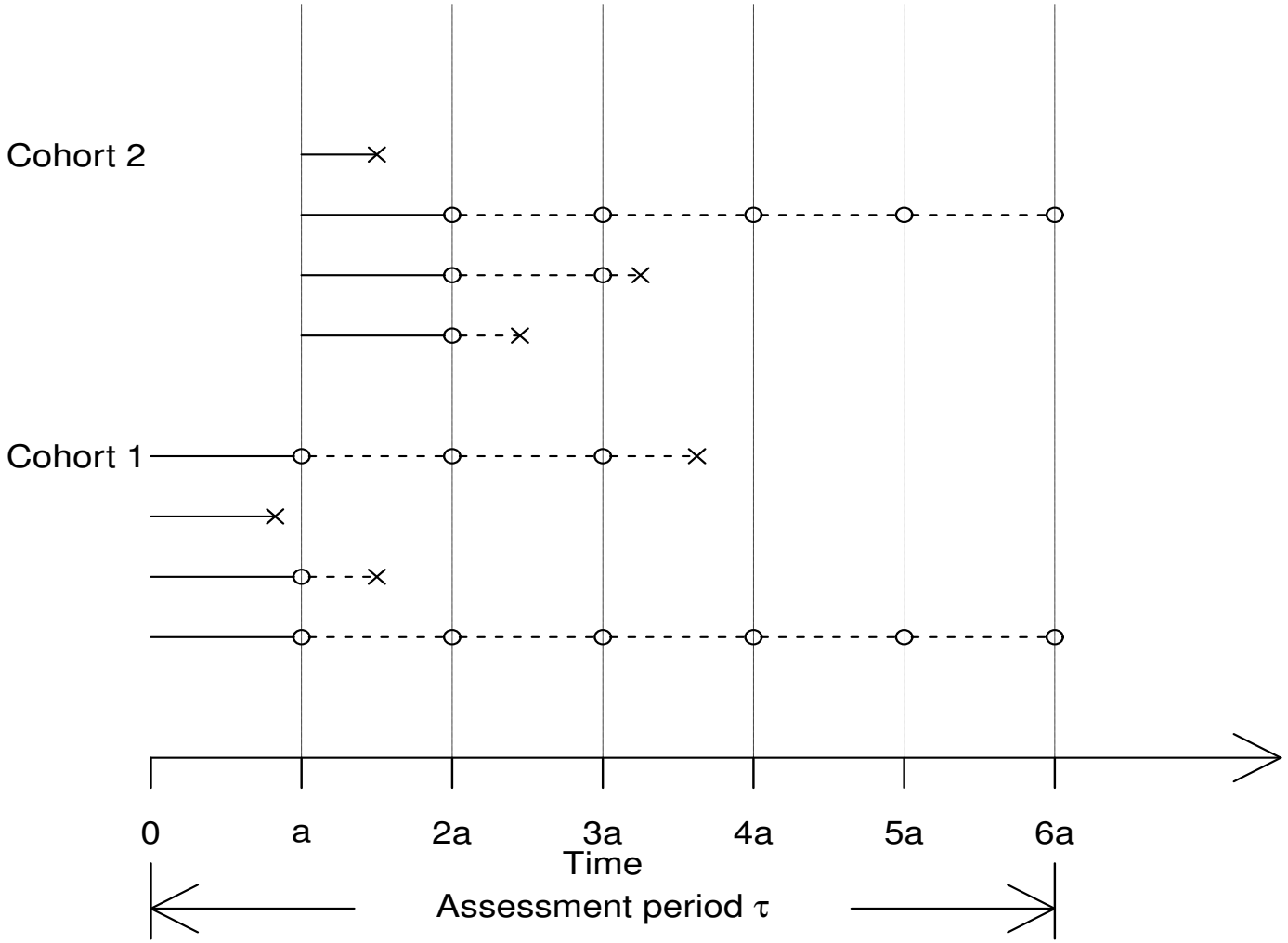


Figure 3: By the time a new cohort is ready for treatment, some of the patients in the trial may be partially followed and their efficacy outcomes have not yet been observed.

Parametric Survival Model

- **Zhang and Rosenberger (2007) developed an optimal allocation scheme under the assumption of parametric survival models.**
- **Let T denote the survival time; under an exponential model the survival function of T is given by**

$$S_j(t) = \exp(-\lambda_j t) = \exp\left(-\frac{t}{\theta_j}\right), \quad j = 1, 2,$$

  where λ_j is the constant hazard rate for treatment arm j, and $\theta_j = 1/\lambda_j$ is the mean survival time.
- **Let Δ_{1i} and Δ_{2i} be the censoring indicators in group 1 and group 2, respectively. Denote $\delta_1 = E(\Delta_{1i})$ and $\delta_2 = E(\Delta_{2i})$.**

Time-to-Event Endpoint

- **Consider the hypothesis test**

$$H_0\text{: } \theta_1 = \theta_2 \quad \text{versus} \quad H_1\text{: } \theta_1 \neq \theta_2.$$

- **The variance of $\hat{\theta}_1 - \hat{\theta}_2$ is**

$$\text{Var}(\hat{\theta}_1 - \hat{\theta}_2) = \frac{\theta_1^2}{n_1\delta_1} + \frac{\theta_2^2}{n_2\delta_2}.$$

- **Zhang and Rosenberger (2007) obtained the optimal allocation ratio by minimizing the total expected hazard $n_1\theta_1^{-1} + n_2\theta_2^{-1}$, subject to fixing the variance as a constant,**

$$r_\theta = \frac{\sqrt{\theta_1^3\delta_2}}{\sqrt{\theta_2^3\delta_1}}.$$

Allocation Ratio with Survival Function

- If the patient response is a good event, then the sooner patients experience the event, the better.
- We minimize the total number of patients who have not responded within the assessment window $(0,\ \tau)$.
- We derive the optimal allocation ratio by minimizing

$$n_1 S_1(\tau,\ \lambda_1) + n_2 S_2(\tau,\ \lambda_2)$$

  subject to fixing $\mathbf{Var}\{S_1(\tau,\ \hat{\lambda}_1) - S_2(\tau,\ \hat{\lambda}_2)\} = K$.
- The optimal allocation ratio is

$$r_S = \frac{\lambda_1 \sqrt{\delta_2 \exp(-\lambda_1 \tau)}}{\lambda_2 \sqrt{\delta_1 \exp(-\lambda_2 \tau)}}.$$

- When the sample size is large and both p_1 and p_2 are small, r_S reduces to that of the binary case, i.e., $r_S \approx \sqrt{p_1}/\sqrt{p_2}$.

Two-stage Response-Adaptive Randomization

- We consider a two-arm clinical trial with binary endpoints,

$$Y_{1i} \sim \textbf{Bernoulli}(p_1), \quad i = 1, \ldots, n_1,$$

$$Y_{2i} \sim \textbf{Bernoulli}(p_2), \quad i = 1, \ldots, n_2.$$

- The null and alternative hypotheses are formulated as

$$H_0 : p_1 = p_2 \qquad \textbf{versus} \qquad H_1 : p_1 \neq p_2.$$

- The trial starts with ER, and continuously makes decisions on when to switch to AR as more data are collected.

Likelihood Ratio Test

- **With m patients in each arm, the likelihood ratio test statistic is**

$$T_m = -2\log\left\{\frac{\max\limits_{H_0:p_1=p_2=p} p^{\sum_{i=1}^m (y_{1i}+y_{2i})}(1-p)^{\sum_{i=1}^m (2-y_{1i}-y_{2i})}}{\max\limits_{p_1,p_2} p_1^{\sum_{i=1}^m y_{1i}}(1-p_1)^{\sum_{i=1}^m (1-y_{1i})} p_2^{\sum_{i=1}^m y_{2i}}(1-p_2)^{\sum_{i=1}^m (1-y_{2i})}}\right\}.$$

- **Under the null hypothesis, the likelihood ratio test statistic follows a chi-squared distribution with one degree of freedom, i.e., $T_m \sim \chi^2_{(1)}$.**
- **We can compute $\hat{T}_m$ by plugging in the MLEs of p_1 and p_2, and the "rejection region" is defined as $\hat{T}_m > \chi^2_{(1)}(1-\tilde{\alpha})$.**

Role of $\tilde{\alpha}$

- As a threshold level for switching from ER to AR, $\tilde{\alpha}$ should be greater than the trial's type I error rate α.
- If the treatment difference is large, n_E would be small so that the trial moves to AR quickly; and if the treatment difference is small, n_E would be large as ER and AR are not much different so that it would take a longer time before switching to AR.
- By controlling $\tilde{\alpha}$, the two-stage design can automatically adapt to the true difference between p_1 and p_2.
- In contrast, if we fix the sample size n_E in the ER stage, it would not be adjustable to the treatment difference.

Two-stage Procedure

- In stage 1, the trial begins with equal randomization, and continuously updates the likelihood ratio test statistic after enrolling every new patient. If $\hat{T}_m < \chi^2_{(1)}(1-\tilde{\alpha})$, equal randomization remains; otherwise, the trial proceeds to stage 2.
- In stage 2, we start to implement response-adaptive randomization for each patient based on an optimal allocation ratio, e.g., using $\sqrt{p_1}/\sqrt{p_2}$ as the allocation ratio to minimize the number of nonresponders.

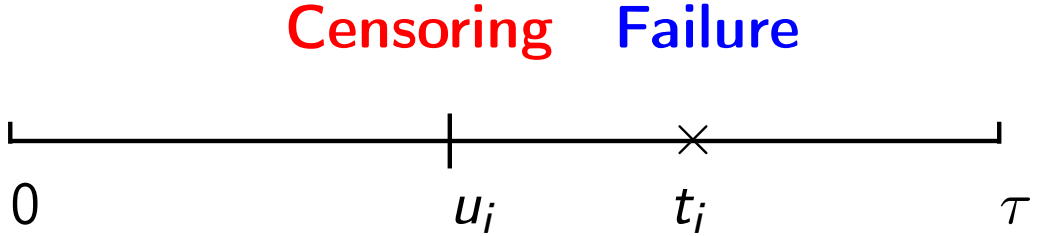
Nonparametric Fractional Model for Delayed Response

- The missing or censoring of response poses immense difficulties when applying response-adaptive randomization during the trial conduct.
- If we view the efficacy endpoint as an event of interest, we can model the time to efficacy using the Kaplan-Meier estimator, and fractionize the censored observations based on patients' exposure times in the trial.
- If a drug-related efficacy event occurs, it is expected to occur within the observation window $[0, \tau]$.

$$Y = \begin{cases} 0 & \text{if the subject does not respond within } [0, \tau], \\ 1 & \text{if the subject responded within } [0, \tau]. \end{cases}$$

Fractional Contribution for Censored Data

- Let T_{1i} denote the time to efficacy, and let u_{1i} $(u_{1i} \leq \tau)$ denote the actual follow-up time for subject i in arm 1.
- The patient's response is censored if he/she has not responded $(u_{1i} < T_{1i})$ and also has not been fully followed up to τ $(u_{1i} < \tau)$.
- If we observe a censored event before τ, i.e., efficacy has not occurred yet, we can obtain a fraction of 1 as the contribution of the censored observation to the response probability.



Redistribution to the Right (Kaplan–Meier Estimator)

- If subject i is censored by the decision-making time u_{1i}, we take the fractional contribution as

$$\mathbf{Pr}(T_{1i} < \tau | T_{1i} > u_{1i}) = \frac{\mathbf{Pr}(u_{1i} < T_{1i} < \tau)}{\mathbf{Pr}(T_{1i} > u_{1i})}.$$

- A fractional contribution for a censored observation is

$$\hat{y}_{1i} = \frac{\hat{S}_1(u_{1i}) - \hat{S}_1(\tau)}{\hat{S}_1(u_{1i})},$$

  where $\hat{S}_1(\cdot)$ is the Kaplan–Meier estimator for arm 1.
- The estimated response rate is $\hat{p}_1 = \sum_{i=1}^{n_1} r_{1i}/n_1$, where

$$r_{1i} = \begin{cases} 0 & \text{if patient } i \text{ does not respond,} \\ 1 & \text{if patient } i \text{ has responded,} \\ \hat{y}_{1i} & \text{if the response of patient } i \text{ is censored.} \end{cases}$$

Simulations

- Our simulation studies considered a two-arm trial with binary outcomes for investigating the operating characteristics of the proposed two-stage fractional AR design.
- The assessment period for efficacy was $\tau = 12$ weeks.
- The accrual time interval between two consecutive cohorts was $a = 1$ week, i.e., every week a new cohort (4 patients) would enter the trial.
- The sample size was calculated based on the type I error and type II error rates, $\alpha = 0.1$ and $\beta = 0.2$ for a two-sided test.
- We fixed the threshold level for ER/AR switching $\tilde{\alpha} = 0.3$.
- For each configuration, we replicated 10,000 trials.

Weibull Distributions

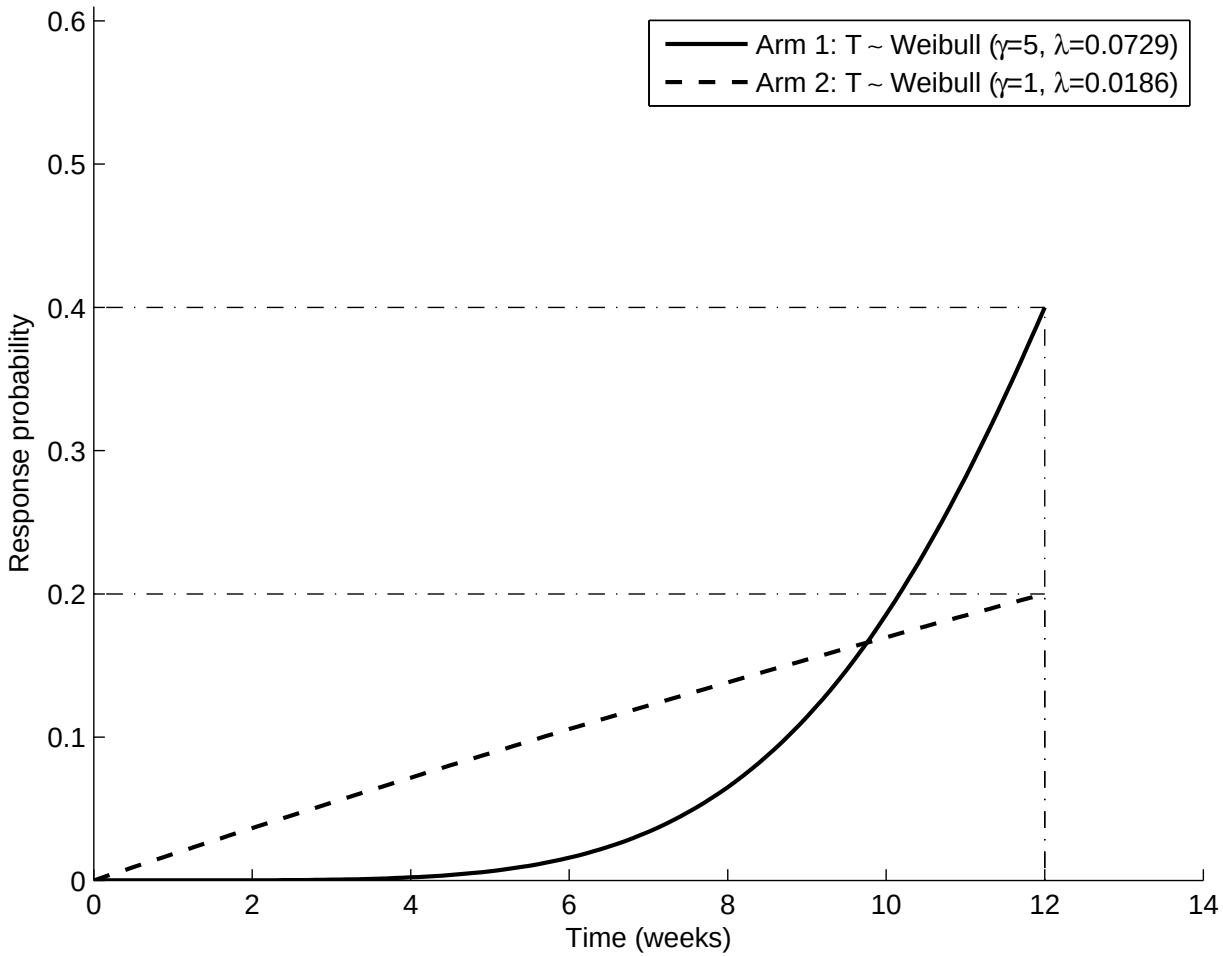


Figure 4: Weibull CDFs with the response probability at time τ being 0.4 for arm 1 and 0.2 for arm 2. The response probability of arm 2 is clearly higher than that of arm 1 before week 10.

Comparison of Three Methods

- Complete-data AR follows each subject till the occurrence of response or the end of the assessment period prior to randomizing each new patient.
- Fractional AR utilizes the scheme of redistribution to the right for censored data, so that each patient would be immediately randomized upon arrival.
- Observed-data AR is based on the observed efficacy data only, while treating censored patients (who have not responded or have been fully followed yet) as nonresponders.

Comparison of AR Designs

Table 1: Comparison of the two-stage observed-data, complete-data, and fractional AR designs with $p_2 = 0.2$ and $n = 132$.

p_1	Two-stage design	Allocation arm 1 (%)	Allocation S.D.	Number of responders	Rejection rate (%)	Trial duration	ER n_E
0.2	Observed	47.4	0.06	26.5	10.1	52.7	56.9
	Complete	50.0	0.06	26.4	10.5	362.1	53.9
	Fractional	50.3	0.07	26.4	9.8	52.7	45.2
0.4	Observed	53.5	0.06	40.5	81.1	53.4	44.5
	Complete	57.8	0.06	41.6	80.5	370.8	27.5
	Fractional	57.5	0.07	41.6	80.9	53.4	30.1
0.6	Observed	57.4	0.06	56.7	99.9	53.6	30.8
	Complete	62.5	0.06	59.3	99.9	372.7	16.1
	Fractional	61.8	0.06	59.0	99.9	53.6	22.1
0.8	Observed	60.7	0.06	74.4	100.0	53.5	22.4
	Complete	65.7	0.05	78.4	100.0	370.0	12.5
	Fractional	64.8	0.06	77.7	100.0	53.5	17.7

Simulation Results

- For $p_1 = 0.2$, the two treatments have the same response rate, all three designs maintained the type I error rate at $\alpha = 0.1$.
- Since a much higher response rate in arm 2 was observed at the beginning of the follow-up, the observed-data AR design falsely assigned more patients to arm 2.
- For $p_1 = 0.4$, it corresponds to the alternative hypothesis, which thus has the targeting power of 80% under all three designs.
- The fractional and the complete-data designs yielded similar allocation ratios, while both are better than the observed-data design.

Simulation Results

- As the difference between the two response rates increases, the sample size of ER becomes smaller because fewer patients are needed to detect a larger difference.
- For $p_1 = 0.8$, fractional AR increased the number of responders by more than three patients over the observed-data design.
- Comparing the duration of the trial between the proposed fractional design and the complete-data design, the trial time was dramatically reduced from 370 weeks to 53 weeks.

Discussion

- The two-stage fractional design addresses two practical issues for response-adaptive randomization:
  (a) **the number of patients in the ER stage is not clearly defined,**
  (b) **patient response cannot be observed quickly enough for real-time AR.**
- In the new design, unobserved efficacy outcomes are naturally treated as censored data, and their fractional point masses are calculated to help making decisions on treatment assignment.
- The nonparametric fractional design is robust and easy to implement, as it only uses the Kaplan–Meier estimator.
- The likelihood ratio test with $\tilde{\alpha}$ is only used for deciding when to switch from ER to AR.

Main References

- Xu, J. and Yin, G. (2014). Two-stage adaptive randomization for delayed response in clinical trials. *Journal of Royal Statistical Society C–Applied Statistics*, 63, 559–578.
- Yin, G. (2012). *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*, John Wiley & Sons (Wiley Series in Probability and Statistics).
- Berry, D. A. and Eick, S. G. (1995). Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine*, 14, 231–246.
- Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 831–853. University of California Press, Berkeley.
- Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N. and Ricks, M. L. (2001). Optimal adaptive designs for binary response trials. *Biometrics*, 57, 909–913.
- Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64, 131–146.

Questions?


תודה
Dankie
Gracias
Спасибо
شكراً
Merci
Takk
Köszönjük
Terima kasih
Grazie
Dziękujemy
Dėkojame
Ďakujeme
Vielen Dank
Paldies
Kiitos
Täname teid
谢谢
Thank You
Tak
感謝您
Obrigado
Teşekkür Ederiz
감사합니다
Σας ευχαριστούμε
ขอบคุณ
Bedankt
Děkujeme vám
ありがとうございます
Tack