Seamless Phase IIa/IIb and Enhanced Dose Finding Adaptive Design

Jiacheng Yuan[a], Herbert Pang[b,c], Tiejun Tong[d], Dong Xi[e], Wenzhao Guo[e], Peter Mesenbrink[e]

[a] Bayer HealthCare Pharmaceuticals Inc., Whippany, New Jersey, USA

[b] Department of Biostatistics and Bioinformatics, Duke School of Medicine, Durham, North Carolina, USA

[c] School of Public Health, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, China

[d] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

[e] Novartis Pharmaceuticals Corporation, One Health Plaza, East Hanover, New Jersey, USA

Received September 14, 2015; Accepted September 4, 2015
Address correspondence to Jiacheng Yuan, Bayer HealthCare Pharmaceuticals Inc., 100 Bayer Boulevard, Whippany, NJ 07981, USA. E-mail: jason.yuan@bayer.com

Abstract

In drug development, when the drug class has a relatively well defined path to regulatory approval and the enrollment is slow with certain patient populations, one may want to consider combining studies of different phases. This paper considers combining a proof of concept (POC) study and a dose finding (DF) study with a control treatment. Conventional DF study designs sometimes are not efficient, or do not have a high probability to find the optimal dose(s) for phase III trials. This paper seeks more efficient DF strategies that allow economical testing of more doses. Hypothetical examples are simulated to compare the proposed adaptive design versus the conventional design based on different models of the overall quantitative representation of efficacy, safety and tolerability. The results show that the proposed adaptive

design tests more active doses with higher power and comparable or smaller sample size in a shorter overall study duration for POC and DF, compared to a conventional design.

Key words: Adaptive design, dose finding, proof of concept, seamless design, utility

1. Introduction

Adaptive designs have often been considered in pharmaceutical industry to improve efficiency, flexibility and the ethics of minimizing patient risks of exposure to ineffective treatments. One of the most common adaptations is a two-stage seamless adaptive design, which combines two separate trials into a single trial. This type of design is especially interesting when the drug class has a relatively well defined path to regulatory approval, and the enrollment is slow with certain patient populations.

Most phase I studies are normally performed in healthy subjects, hence are less suitable for combination with phase II/III patient studies. Exceptions occur in oncology where phase I studies are commonly conducted with patients, see Gooley et al. (1994), Thall and Russel (1998), Thall and Cook (2004), and Hoering et al. (2011). Therefore combination of phase I with phase II into a seamless adaptive design is uncommon in therapeutic areas other than oncology. In the literature, seamless phase II/III designs have been extensively discussed, Bretz et al. (2006) reviewed the general concepts, Schmidli et al. (2006) reviewed various applications, and Stallard and Todd (2010) compared various statistical approaches in a unified framework. However, in many situations (especially in non-oncology therapeutic areas) there is reluctance to adopt seamless II/III design, as normally it is hard for health authorities (HA's) to accept adaptive

designs in pivotal trials, and a gap between phase II and III is also quite often necessary for a sponsor to make decisions as well as to get feedback from HA's.

Phase II itself usually includes multiple studies in different stages, among them are a proof of concept (POC) study (phase IIa) and a dose finding (DF) study (phase IIb). A POC study is usually conducted to demonstrate clinical efficacy with a small number of patients, and the primary endpoint is assessed at a relatively early time point post-dosing. If the POC study provides proof of efficacy, a DF study is often conducted to assess the efficacy as well as safety with a relatively larger number of patients to find the optimal dose(s) for phase III confirmatory trials, where the primary efficacy endpoint is assessed at a later time point post-dosing. Conventionally, DF studies have a small, fixed number (3 or 4) of active doses, and try to choose the most promising dose(s) among them. This approach works well with an important assumption that conventional studies do not always meet: the doses assessed include the optimal dose. The conventional approach also usually powers DF studies at the same level for every dose, which may be unnecessary. The high cost of carrying all arms of a phase II DF trial for the full duration of a study usually limits the number of doses tested. Testing relatively few doses limits the acquisition of knowledge about the test drug, and also increases the likelihood of selecting a dose for confirmatory trials that is not at the optimal therapeutic level. This increases the risk of failure in the current phase or a more costly failure in a subsequent study, and sub-optimal dose selection in phase II studies contributes significantly to the failure in phase III. Thus, it is desirable to seek more efficient DF approaches that allow economical testing of more doses to improve chances of correctly identifying optimal doses for phase III trials (Rosenberg, 2010).

The first component of the proposed adaptive design is the combination of a POC study and a DF study with a control treatment (which can be a placebo or background therapy). The second component of the proposed adaptive design is the enhancement in dose finding. The key features of the proposed design are to start with more dose levels, e.g. 6 or 7 doses, and conduct the trial in two stages, where in the first stage establish proof of efficacy via comparison of active drug to placebo using pooled groups excluding the worst performers, then recycle the majority of placebo-treated patients into the second stage. In the second stage of dose finding, drop less promising doses sequentially, and in the end test the superiority between the remaining individual active dose(s) and placebo.

In real situations at trial design stage, sample size or power is usually calculated based on historical assumptions of efficacy endpoint(s), and safety is not considered because normally it is not known and hard to make an appropriate quantitative assumption. This paper introduces the concept of utility as an overall quantitative representation of efficacy, safety and tolerability, and compares the proposed adaptive design with the conventional design under several postulated utility models. The results show that the proposed adaptive design tests more active doses with higher power and comparable or smaller sample size in a shorter overall study duration for POC and DF, compared to a conventional design.

The rest of the paper is organized as follows. The concept of utility is introduced in Section 2, and the adaptive design with seamless phase IIa/IIb transition and enhanced dose finding approach is described in Section 3. In Section 4, the proposed adaptive design is compared to the

conventional design with simulations under several utility models. The paper concludes in Section 5 with some discussions.

2. Utility

In normal situations, there is a minimum effective dose level, beyond which the therapeutic effect (desired drug effect) starts to show, then the therapeutic effect rises with the increase of dose levels, until a certain point after which the therapeutic effect reaches a plateau despite the increase of dose levels. In addition to the desired drug effect, there are also side effects, which usually occur with the increase of dose levels. Therefore, a higher dose with worse side effects can in general make it less desirable. Utility is used to account for both therapeutic effects (efficacy) and side effects (safety and tolerability). Usually the sample size or power calculation is based on efficacy endpoint(s) only. When safety and tolerability are also taken into consideration, the actual power may be lower than the nominal level as planned. In the simulations that follow, utility is used to compare the power between the suggested adaptive design and the conventional design, where utility is an overall quantitative representation of efficacy, safety and tolerability.

In the following example, the mean values of different doses are hypothetical, i.e. not actual values from studies, to show a postulated response model. The sample size of 50 per arm is an arbitrarily picked number. The standard deviations (SDs) and residual values from ANOVA models are from a psoriatic arthritis (PsA) study of Novartis, but with disguise (i.e. with addition of non-zero constant values) to keep confidentiality of sensitive information.

Assume in a PsA study percent improvement in American College of Rheumatology response (ACRn) is assessed for efficacy. ACRn is a continuous endpoint with higher value indicating better efficacy. Utility is defined as follows,

$$U_{PsA} = ACRn - 10 \times I[AE] - 20 \times I[SAE],$$

where I[AE] and I[SAE] are indicator functions taking value of 1 if an event happens and 0 otherwise. If a patient has an AE but not an SAE, then ACRn will be penalized by 10 for the utility, and for a patient with an SAE the ACRn will be penalized by 30 for the utility. Table 1 shows the summary of ACRn and utility results.

In the above example utility is a linear function of a continuous efficacy endpoint and occurrence of AE and SAEs. In this paper, the interest is not to estimate the utility dose-response curve or establish a mathematical model. Utility is generalized to be a function of multiple discrete and/or continuous endpoints, and the utility function may or may not have an explicit mathematic formula. The situation without a mathematic formula is actually common and is routinely applied in practice – the dose is chosen based on overall assessment of efficacy and safety results, although the primary variable(s) is most often an efficacy endpoint(s).

In the simulations presented in Section 4, there is no specific formula for the utility function. Although the utility function is unknown, in a wide range of postulated models, it is shown that the proposed design has better performance. This is similar to what is often seen in the existing literature.

Depending on how the desired drug effects and side effects are impacted by the increase of dose levels, there can be different utility models. Table 2 shows 8 utility models that are considered in simulations. For each model, two sets of utility numbers across 6 active doses are displayed, one set for POC stage and the other for DF stage.

The utility dose-response curves at the DF stage of the 8 models are shown in Figure 1, where the numbers indicate the indices of the corresponding models. The basic principles for these models are as follows:

- The best dose at both the POC and the DF stage (may or may not be the same dose level) has a utility of 100.

- For POC, as it is an earlier time point, safety and tolerability issues would not negate much of the benefit, the worst dose at POC stage has a higher utility than at the DF stage.

- The placebo is assumed to have a utility of 20 at POC time point, and 15 at DF time point.

- SD, which is 126 for the POC utilities and 164 for the DF utilities, is specified such that the power based on utility for the best dose is 95% of the nominal power for POC and 90% of the nominal power for DF with the conventional approach.

One could deem these models are based on normalized parameters from the hypothetical examples with assumed dose-response trajectories.

Our proposed design is not limited to a specific indication, but non-oncology therapies in general. Utility function would be different for different indications. When the design is used in a specific setting, the utility function would better be established with agreement from clinicians and HA's.

The utility doesn't have to be normally distributed, as there are corresponding statistical methods in existence. The assumption of normality in the example in Section 4.2 is only for ease to compare the operating characteristics. As shown later, the utility and the original efficacy endpoint ACRn are similarly 'non-normal'. In practice, it could be very well accepted for ACRn to be analyzed as a normal variable. (If one is prudent enough to use non parametric method then the same can be used for utility.) One may prefer the utility to be unit free, that is the rationale for the generalization of utility. On the other hand, for a specific indication in application, this should be an easy thing to handle, just a matter of transformation.

3. Seamless phase IIa/IIb transition and enhanced dose finding approach

Similar to Chow et al. (2007) and Chow and Tu (2008), we assume the endpoints for phase IIa and IIb studies are similar but different, e.g. a biomarker versus a clinical endpoint or the same study endpoint evaluated at different time points. The primary endpoint is denoted for the two stages with E_{poc} and E_{df}, respectively, which are correlated. Often two active dose levels are kept in the phase III studies, and this is the scenario that will be focused on here.

To improve chances of correctly identifying optimal doses for phase III trials, m active dose levels are initiated which are believed to include the optimal dose with a certain level of confidence. Practically, m can be 6 or 7, which should be a reasonable number given that toxicity and phase I research have already finished, otherwise the number of doses may increase if there are good reasons. For each dose, the corresponding group has two subgroups, immediate-treatment and delayed-treatment, where the half in immediate-treatment subgroup are dosed with active treatment at the beginning, while the other half in delayed-treatment subgroup are dosed with placebo at the beginning. One group can enroll up to G patients at the maximum, and G is the per arm sample size that is adopted by a conventional DF study.

The trial is conducted in two stages, where in the first stage proof of efficacy is established via comparison of active drug to placebo using pooled groups excluding the worst performers, then the majority of placebo-treated patients are recycled into the second stage. In the second stage of dose finding, less promising doses (as decided by descriptive statistics, e.g. mean of utility) are dropped sequentially. In the end superiority is tested between the remaining individual active dose(s) and placebo. Let $\mu_{a,poc}$, $\mu_{p,poc}$ be the mean of utility for pooled active doses and placebo at POC stage, and $\mu_{a,df}$, $\mu_{p,df}$ the mean of the optimal dose and placebo at DF stage, then the tests for POC and DF are as follows.

$$H_{10}: \mu_{a,poc} - \mu_{p,poc} \leq 0 \text{ vs. } H_{1a}: \mu_{a,poc} - \mu_{p,poc} > 0 \quad \text{(i)}$$

$$H_{20}: \mu_{a,df} - \mu_{p,df} \leq 0 \text{ vs. } H_{2a}: \mu_{a,df} - \mu_{p,df} > 0 \quad \text{(ii)}$$

Figures 2 and 3 show the study process and flowchart of the adaptive design.

Let G_0 and G denote per arm sample size for the POC and DF trials, respectively, in a conventional design, where POC has 1 active arm and 1 placebo arm, and DF has 3 active arms and 1 placebo arm. For an example where m = 6, B = 4, $\gamma = 0.2$, with the suggested adaptive design the sample size is 4.2G, while the sample size of the conventional design is $(2G_0 + 4G)$.

4. Operating characteristics

In this section we compare operating characteristics of the proposed adaptive design with 210 patients testing 6 doses, versus a POC study with 74 patients and a DF study with 200 patients under the conventional design that tests 4 doses (or 3 doses if the POC dose is one of those in DF). Model 1 of the utility distribution is used as an example to explain in details how the simulation is conducted, but results for all 8 models are presented in the end.

4.1. Correlations

In the discussed setting with comparisons of multiple doses versus placebo at more than one time points, there are two types of correlations. One is the structural correlation of the test statistics between doses, because of a common placebo as the comparator. The other is the temporal correlation because some patients contribute to both the POC test and the DF test. The structural correlation is accounted in Section 4.2, and the temporal correlation is accounted in Section 4.3.

4.2. Conventional approach

Using the PsA example, if in the future a compound of the same class is to be developed, information from this trial can be used as part of the assumptions needed for calculating the sample size. As U_{PsA} = ACRn – 10×I[AE] – 20×I[SAE], where ACRn is usually assumed to be normally distributed, I[AE] and I[SAE] are two binary variables, one would not think U_{PsA} is still normally distributed. However, in this example, probably due to relatively few AEs and SAEs as well as the fact that ACRn is a composite endpoint with 7 components, the utility is still approximately a normal random variable. Or in other words, U_{PsA} does not seem to be more distant from normal distribution than ACRn. Figure 4 shows the histogram of the residuals from ANOVA models with ACRn and U_{PsA}, respectively, as the dependent variable and treatment as the explanatory variable.

In this example, utility is a linear function of a continuous efficacy endpoint and occurrence of AE and SAE, and it is assumed to follow normal distribution. The method can be generalized to situations where the utility is a function of a discrete efficacy endpoint and other safety endpoints, takes continuous values, but is not a normally distributed variable, in which case, the hypothesis testing is conducted with different procedures, e.g. using Wilcoxon rank sum test instead of t-test.

In a conventional design, the POC study that assesses the ACRn at Week 4 as the primary endpoint, with 1 active arm and 1 placebo arm would require 74 patients, i.e. 37 per arm, to achieve 80% power (assuming the mean is 16.0 and -9.5, respectively, for the active and placebo arm, with a common standard deviation of 38.5, and a significance level of 0.05 two sided). For

the DF study in a conventional style, 3 active arms and 1 placebo arm with ACRn at Week 12 as the primary endpoint, without adjustment of multiplicity, the per arm size will be 50 patients, to achieve 80% power (assuming the mean is 25.3 in an active arm, and -3.8 in the placebo arm, with a common standard deviation of 51.2, and a significance level of 0.05 two sided).

4.2.1. POC power in terms of utility

In terms of the utility specified in Table 1, if still assume normal distribution for utility, then the power for the POC (test at Week 4) drops to 76%, assuming the mean is 14.0 and -14.5, respectively, for the active and placebo arm, with a 1:1 randomization, a common standard deviation of 45, a significance level of 0.05 two sided, and a per arm sample size of 37.

Using numbers from Model 1 of utility in Table 2, the power of 76% for the POC study is derived with the active mean of 100, the placebo mean of 20, a common standard deviation of 126, assuming a 1:1 randomization, a significance level of 0.05 two sided, and a per arm sample size of 37.

4.2.2. DF power in terms of utility

In terms of the utility specified in Table 1, for the DF (test at Week 12), if only consider one active dose compared to placebo as has been a routine in practice, the power is 72%, assuming the mean is 20.0 in an active arm, and -6.8 in the placebo arm, with a 1:1:1:1 randomization, a common standard deviation of 52, a significance level of 0.05 two sided, and a per arm sample size of 50. Using numbers from Model 1 of utility in Table 2, the power of 72% for the DF study is derived with the active mean of 100, the placebo mean of 15, a common standard deviation of

164, assuming a 1:1:1:1 randomization, a significance level of 0.05 two sided, and a per arm sample size of 50. However, as to be discussed below, when there are multiple comparisons, it is more complicated.

In practice, two testing procedures are not uncommon in DF studies with conventional approach. One is to test the hypothesis for each dose the same way as if there were only one dose, another is to test these doses in a pre-specified sequence, e.g. from high dose to low dose (lower dose is tested only when the null hypothesis has been rejected for higher ones). The former does not adjust for multiplicity, but the latter does and controls the family-wise type I error. The R package gMCP is used to assess the operating characteristics for the conventional approaches, see Bretz et al. (2009, 2011a, 2011b). The two testing procedures are illustrated in Figure 5, where for Strategy 1 on the top each hypothesis is tested at α level without propagation, and for Strategy 2 at the bottom H_1 is tested first at α level (0 for H_2 and H_3 means they are not tested at the beginning) and H_2 will be tested if and only if H_1 is rejected, and the same from H_2 to H_3

Assuming the conventional approach randomly picks 3 out of the 6 doses to conduct the DF study, gMCP is used to calculate the following operating characteristics:

(1) Power: the probability that the highest-utility dose (i.e. the optimal dose) is included in the DF study and its superiority over placebo is established, under the condition that the utility of the test drug is greater than placebo (at all dose levels).

(2) Type I error rate: the probability that superiority over placebo is established for at least one dose, under the condition that the utility of the test drug is the same as the placebo (at all dose levels).

Let H1, H2, H3 represent the null hypotheses associated with chosen doses from high to low, the transition matrices for the two strategies are as follows:

$$G_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad G_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

All elements in G_1 are 0 because there is no alpha propagation in this procedure. In G_2, the full alpha is propagated from H_1 to H_2, and again from H_2 to H_3, and nowhere else, therefore 1 in the cells (1, 2) and (2, 3), and 0 in all other cells.

Let the following matrix R represent the correlation coefficients between test statistics of each two out of the three hypotheses.

$$R = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}.$$

As the structural correlation induced by comparing two doses with a common control in a balanced parallel group design is 0.5, see Bretz et al. (2011b), simulations are performed for ρ = 0.5, 0.7, 0.9, 1, and the results are shown in Table 3. With the increase of correlation between doses, power is increased under Strategy 2 and type I error rate is decreased under Strategy 1,

such that when the doses are independent power is the lowest (under Strategy 2) and type I error rate is the highest (under Strategy 1).

In theory, it is possible that the three doses assessed include the optimal one, and that all three doses are superior to placebo; however, the optimal dose is not chosen for phase III (i.e. the other two doses perform better in the trial). Therefore, the power shown in Table 4 could be higher than the preferred situation – choosing the optimal dose for phase III. Given that the goal is to show that the proposed adaptive design has higher power than the conventional approach, this definition is used for the conventional approach (which allows an 'inflated' power).

4.2.3. Overall operating characteristics for conventional approach

With the conventional approach, DF power is a conditional probability conditioned on a successful POC study. Therefore, the power of winning both is the product of the POC power and the DF power. Under Model 1 utility distribution, (ignoring the alpha inflation with Strategy 1), the power for the conventional approach is in the range of 0.16 - 0.25 for winning both POC and DF.

4.3. Proposed adaptive design

As shown in previous sections, when the results of different doses are independent (i.e. the structural correlation is 0.5), the type I error rate (for Strategy 1) and power (for Strategy 2) are the worst with the conventional approach. It should be similar for the adaptive design, i.e. power and type I error are the worst when the results of different doses are independent. In what follows, the situation is assessed where the results of different doses are independent.

Among the patients that contribute to the active treatment of the final DF test, 20% that are on active treatment from the beginning also contribute to POC test for the active treatment. Likewise, among the patients that contribute to placebo treatment of the final DF test, 20% that don't have a dosing restart also contribute to POC test for the placebo treatment. The simulations account for the correlation between the POC and DF endpoints, i.e. the temporal correlation, for these patients. (There could also be patients who contribute to placebo at POC but active treatment at DF. These correlations are ignored.) Different levels of temporal correlation are evaluated. For each level of the correlation within the set of (0, 0.5, 0.7, 0.9, 1), 10,000 simulation runs are performed to calculate the following operating characteristics:

(1) Power: under the condition that the utility of the test drug is greater than the placebo (at all dose levels), the proportion of successes with superiority established for both POC and DF, and the optimal dose is in the two doses picked for phase III

(2) Type I error rate: under the condition that the utility of the test drug is the same as the placebo (at all dose levels), the proportion of successes with superiority established for both POC and DF

With the Model 1 utility distribution, the results are shown in Table 4. The impact of correlation is negligible. (The situation is similar with other models.)

The type I error rate is well controlled for the proposed adaptive design. The power is in the range of 0.31- 0.32 for winning both POC and DF and the two doses selected for phase III include the optimal dose.

4.4. Comparison of power between the adaptive design and the conventional design

Under different scenarios of the utility model, the power for the proposed adaptive design and the conventional design is shown in Table 5, where power is defined as follows:

- with the conventional design the probability of (1) a successful POC and (2) the highest-utility dose is included in the DF study and its superiority over placebo is established, under the condition that the utility of the test drug is greater than the placebo.

- with the adaptive design the probability that superiority is established for both POC and DF, and the optimal dose is in the two doses picked for phase III, under the condition that the utility of the test drug is greater than the placebo.

In all the 8 models, the proposed adaptive design has advantage over the conventional design in terms of power. Specifically, in models where doses have relatively larger difference in utilities (Models 2, 4, 7), the advantage is the most pronounced, while in models where doses have similar utilities (Models 5, 8) the advantage is smaller. Other design elements of the conventional approach and the proposed adaptive design are displayed in Table 6. The results from Tables 5 and 6 have shown that the proposed adaptive design has the following advantages compared to a conventional design:

- Tests more active doses

- Higher power

- Comparable/smaller sample size
- Shorter overall study duration for POC and DF

4.5. Control of type I error rate

From the above simulations, in the conventional design, the type I error rate of the DF study may or may not be controlled depending on which testing procedure is adopted. (The POC study controls the type I error rate as there is only one dose.) The adaptive design controls the family-wise type I error rate at 0.003 level.

5. Discussion

5.1. Doses picked with DF assessment

When choosing 2 doses that will be used as the basis for regulatory approval, among several doses where a monotonic dose response relationship exists, e.g. 10 mg, 20 mg, 40 mg, 80 mg, 160 mg, 320 mg, intuitively it makes sense to pick two adjacent doses (as it would be counter-intuitive to approve 20 mg and 80 mg but omit 40 mg). However, since the 2 doses picked from DF will be further evaluated in future phase III (adequate and well controlled) studies, it could be acceptable to pick 2 doses that are not adjacent in the candidate pool of doses. Actually, if the DF study results show that 20mg and 80mg are the best two performers, it would be reasonable that one actually uses 20mg and 40mg (or 40mg and 80mg) for phase III.

5.2. Rationale to have half patients in a delayed-treatment subgroup of every dose level at the beginning

Why not consider several active arms and one placebo arm in parallel with equal size? In the design proposed, a POC test is conducted at an early stage, where the active doses are combined excluding a couple of 'losers', and compared to the placebo group. To have a balance between the pooled active patients and placebo patients, 6 or 7 groups are used at the start of the POC, and each group is balanced between active and placebo (using a stratified randomization). After POC criteria have been met, only two delayed-treatment subgroups remain on placebo, and constitute the placebo group for the DF test later.

5.3. Rationale to perform POC test with better-performing subgroups

If all active doses are included for POC test and ineffective active doses exist, the comparison power will be diluted. Therefore the worst active doses are excluded. The worst placebo subgroups are also excluded to mirror the algorithm for the active subgroups, otherwise the type I error may be inflated because better values are 'picked' for actively-treated patients.

5.4. Rationale for some patients to remain on placebo with a dosing restart in the second stage

The end goal is to perform the final comparison between the placebo group and each of the remaining active groups, where each active group has two types of patients, one type would be

those starting from the very beginning with the active dose, and another type would be those converting to active dose from placebo who have a dosing restart. With the designed handling, the placebo group will also have two similar types of patients, except that they are on placebo.

5.5. Mathematical proof of the type I error control

Formal mathematical proof of type I error control would definitely be valuable in addition to simulations that have been conducted, which would be a topic of interest for further investigation.

Acknowledgement

We thank Dr. Heinz Schmidli for his review and comments on an earlier version of this paper, which have been incorporated into this version. The authors thank the editor, the associate editor, and three referees for their constructive comments that led to a substantial improvement of the paper.

Funding

This work is partially supported by the National Institutes of Health (grant P01CA142538).

References

Bretz, F., Maurer, W., Brannath, W., Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 28: 586–604.

Bretz, F., Maurer, W., Hommel, G. (2011b). Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine* 30: 1489–1501.

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W., Rohmeyer, K. (2011a). Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes or parametric tests. *Biometrical Journal* 53: 894–913.

Bretz, F., Schmidli, H., König, F., Racine, A., Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypothesis selection at interim: General concepts. *Biometrical Journal* 48: 623–634.

Chow, S. C., Lu, Q., Tse, S. K. (2007). Statistical analysis for two-stage seamless design with different study endpoints. *Journal of Biopharmaceutical Statistics* 17: 1163–1176.

Chow, S. C., Tu, Y. H. (2008). On two-stage seamless adaptive design in clinical trials. *Journal of the Formosan Medical Association* 107: 52-60.

Gooley, T. A., Martin, P. J., Fisher, L. D., Pettinger, M. (1994). Simulation as a design tool for phase I/II clinical trials: An example from bone marrowtransplan tation. *Controlled Clinical Trials* 15: 450–462.

Hoering, A., LeBlanc, M., Crowley, J. (2011) Seamless phase I-II trial design for assessing toxicity and efficacy for targeted agents. *Clin Cancer Research* 17: 640–646.Rosenberg, M. J.

(2010). *The Agile Approach to Adaptive Research: Optimizing Efficiency in Clinical Development*. Wiley: Hoboken, New Jersey.

Schmidli, H., Bretz, F., Racine, A., Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim. Part II: Applications and Practical Considerations. *Biometrical Journal* 48: 635–643.

Stallard, N., Todd. S. (2010). Seamless phase II/III designs. *Statistical Methods in Medical Research* 20: 623-634.

Thall, P. F., Russell, K. T. (1998). A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics* 54: 251–264.

Thall, P. F., Cook, J. D. (2004). Dose-finding based on efficacy-toxicity trade offs. *Biometrics* 60: 684–693.

Figure 1. Utility dose-response curves at the DF stage of the 8 models

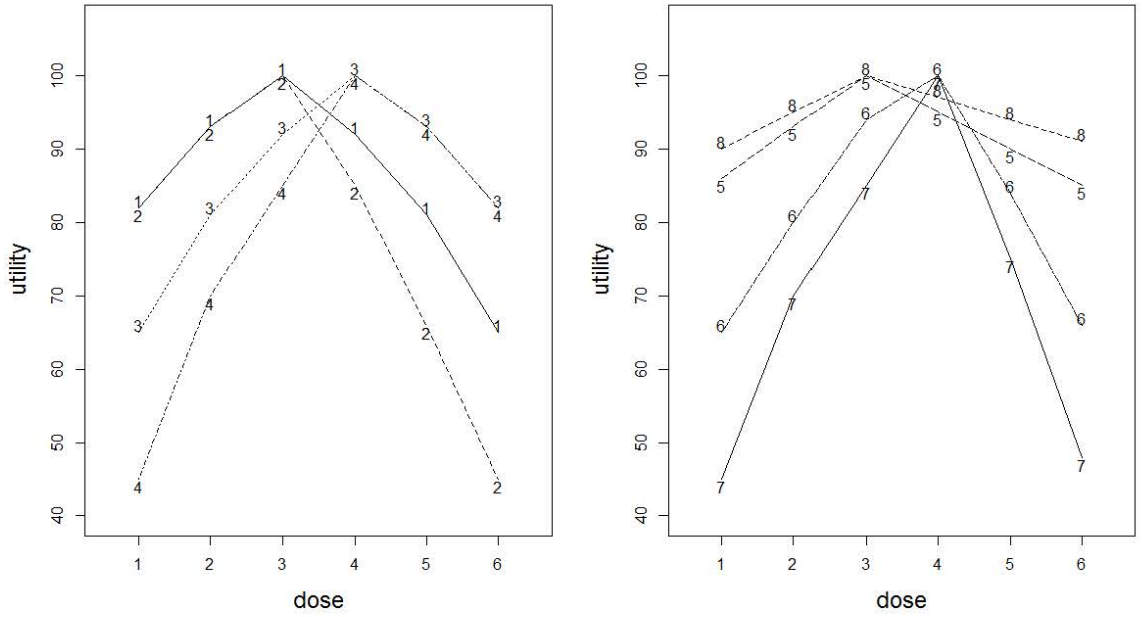
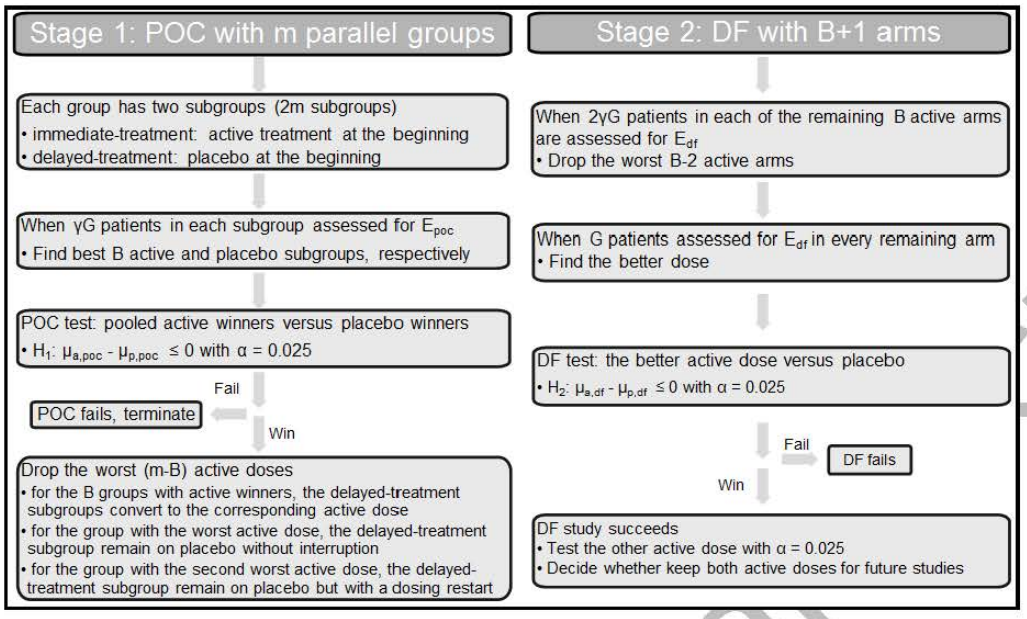
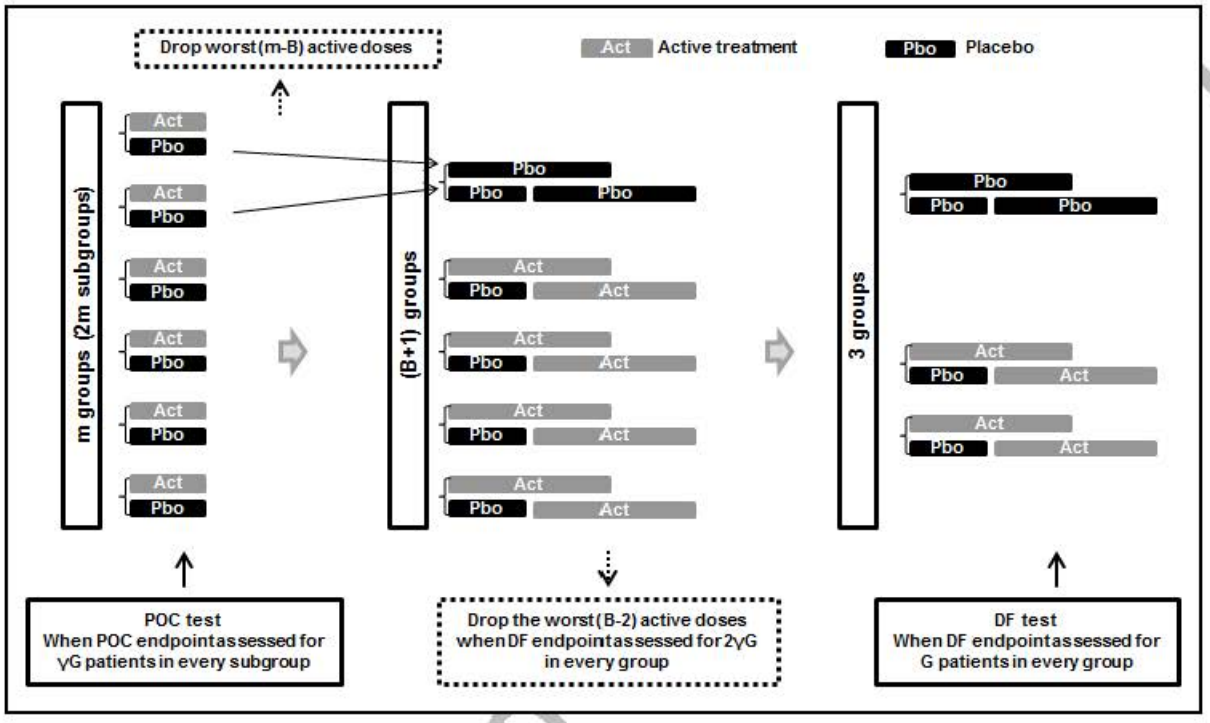


Figure 2. Study process of the adaptive design



The parameter G, per arm sample size adopted by a conventional DF study, is the maximum number of patients that can be enrolled in a group in the adaptive design. Conditions for other parameters are $m \geq 6$, $0 < \gamma < 0.5$, and $3 \leq B \leq m - 2$.**Figure 3.** Study flowchart of the adaptive design



Although there are multiple placebo groups at the beginning, most of them switch to active treatment after POC criteria have been met. The overall exposure time on placebo in this design is comparable or sometimes less than that in a conventional design.

Figure 4. Histogram of residuals from ANOVA model with ACRn and utility, respectively, as the dependent variable, and treatment as the explanatory variable

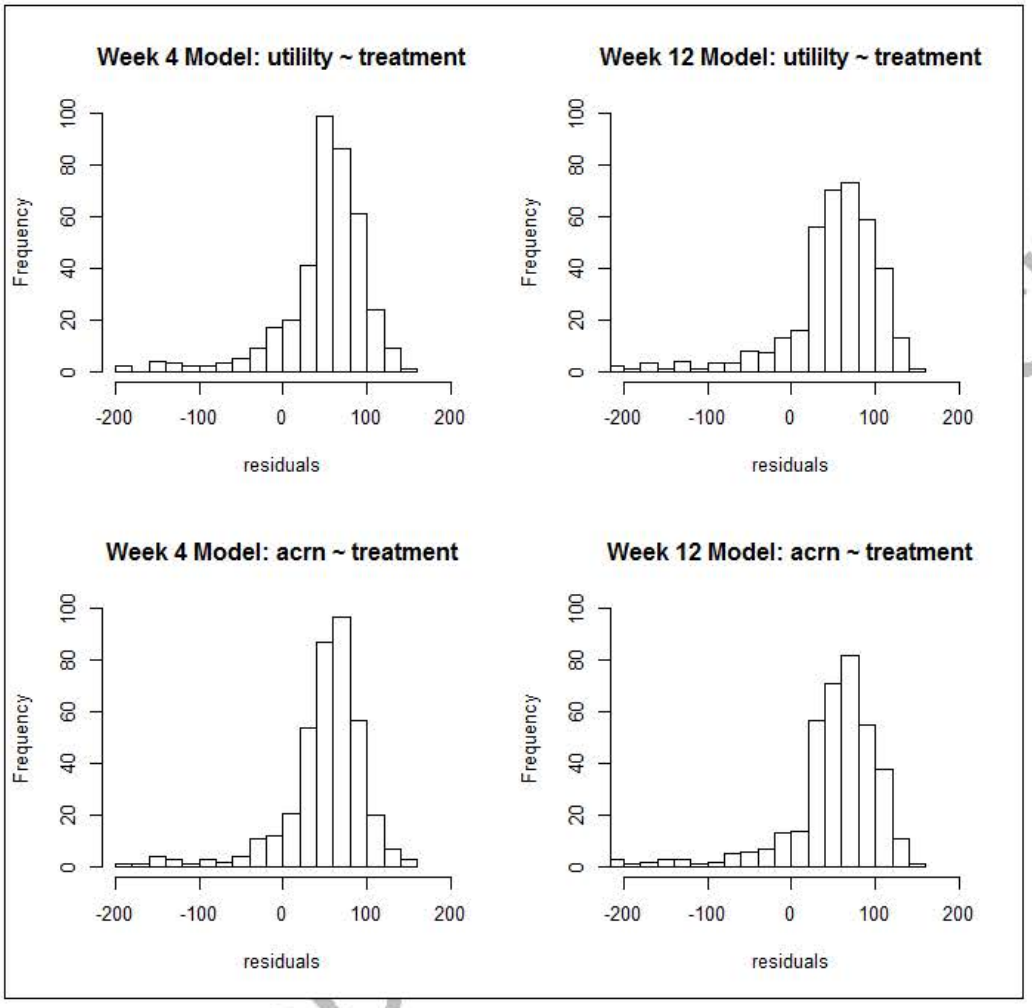


Figure 5. Two strategies in DF studies with conventional approach

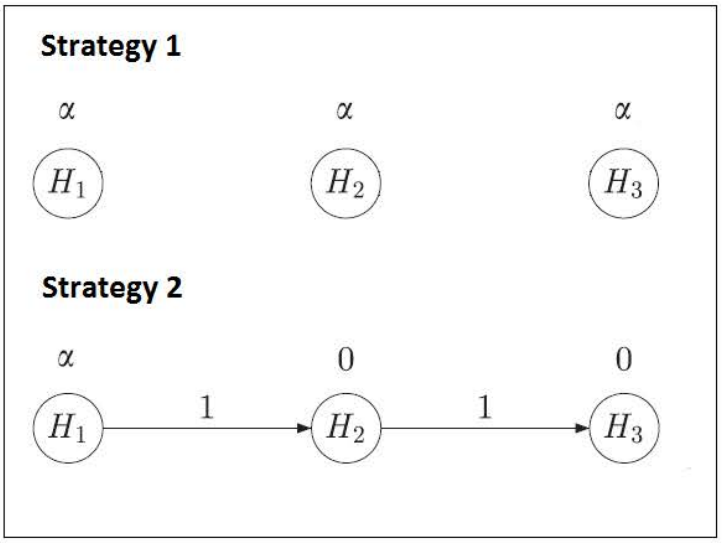


Table 1. Summary of a hypothetical psoriatic arthritis study results

			ACRn		U_{PsA}	
Visit	Dose	n	mean	sd	mean	sd
Week 4	low	50	-2.8	41.6	-5.4	45.3
	medium	50	13.4	43.2	11.3	44.2
	high	50	16.0	37.3	14.0	44.9
	placebo	50	-9.5	39.7	-14.5	45.1
Week 12	low	50	1.4	51.3	-2.3	51.2
	medium	50	22.8	54.5	17.5	54.0
	high	50	25.3	51.9	20.0	52.6
	placebo	50	-3.8	50.5	-6.8	51.3

Table 2. Different utility models considered in simulations

Model	Utility distribution	
	POC	DF
1	90, 92, 97, 100, 92, 80	82, 93, 100, 92, 81, 65
2	90, 92, 97, 100, 85, 60	82, 93, 100, 85, 66, 45
3	80, 92, 100, 97, 92, 90	65, 81, 92, 100, 93, 82
4	60, 85, 100, 97, 92, 90	45, 70, 85, 100, 93, 82
5	88, 95, 100, 97, 92, 87	86, 93, 100, 95, 90, 85
6	80, 92, 100, 95, 89, 81	65, 80, 94, 100, 84, 66
7	75, 89, 100, 94, 86, 77	45, 70, 85, 100, 75, 48
8	92, 97, 100, 98, 96, 93	90, 95, 100, 97, 94, 91

Table 3. Impact of structural correlation on operating characteristics for the two strategies with conventional approaches

Correlation of test statistic between doses	**DF Power under Strategy 2**	**DF Type I error rate under Strategy 1**
0.5	0.21	0.13
0.7	0.22	0.11
0.9	0.24	0.08
1	0.24	0.05

Table 4. Power and type I error rate for adaptive design

Temporal Correlation	Power	Type I error rate
0	0.31	0.001
0.5	0.32	0.002
0.7	0.31	0.002
0.9	0.31	0.002
1	0.32	0.003

Table 5: Power of adaptive design and conventional design

Model	Conventional design	Adaptive design
1	0.16 - 0.25	0.30 - 0.32
2	0.11 - 0.22	0.32 - 0.34
3	0.18 - 0.25	0.31 - 0.32
4	0.16 - 0.22	0.32 - 0.33
5	0.20 - 0.26	0.29 - 0.31
6	0.16 - 0.26	0.31 - 0.33
7	0.12 - 0.22	0.32 - 0.35
8	0.23 - 0.28	0.27 - 0.29

Table 6: Design elements of adaptive design and conventional design

Design elements	**Conventional design**	**Adaptive design**
Total number of patients	274	210
Number of active doses	4 or 3*	6
Patient years of exposure to placebo	14.3	14.6
Gap between POC and DF	Yes	No
**If POC dose is also one in the DF*		