Seamless Phase IIa/IIb and Enhanced Dose Finding Adaptive Design

Jiacheng Yuan[1], Herbert Pang[2], Tiejun Tong[3]
[1]Novartis Pharmaceuticals Corporation, One Health Plaza, East Hanover, NJ 07936
[2]Dept of Biostatistics and Bioinformatics, Duke School of Medicine, Durham, NC 27710
[3]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

Abstract

In drug development, when the product class has a relatively well defined path to market and the enrollment is slow with certain patient populations, one may want to consider combining studies of different phases. This paper considers combination of a proof of concept study and a dose finding (DF) study. Conventional DF study designs sometimes are not efficient, or don't have good chances to find the optimal doses for Phase III trials. This paper seeks more efficient DF approaches that allow economical testing of more doses. Hypothetical examples are simulated to compare the proposed adaptive design versus the conventional design based on assumed models of the overall quantitative representation of efficacy, safety and tolerance. The results show that the proposed adaptive design tests more active doses with higher actual power and comparable or smaller sample size in a shorter overall duration, 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 ethicality. A two-stage seamless adaptive design is one of the mostly considered adaptations (Chow et al, 2007; Chow and Tu, 2008). A two-stage adaptive seamless design combines two separate trials into one single trial, which is especially interesting when the product class has a relatively well defined path to market, e.g. similar drugs have previously been approved, and the enrollment is slow with certain patient populations, e.g. rheumatoid arthritis patients who are anti-TNF inadequate responders. Focusing on combination of Phase II dose finding and Phase III confirmation of efficacy, Chow et al (2007) and Chow and Tu (2008) discussed the analysis based on combined data and derived the sample size calculation formulas, assuming that there is a well established relationship between the endpoints of the two stages. This paper discusses the combination of a proof of concept (POC) study and a dose finding (DF) study, i.e. a seamless Phase IIa/IIb design.

A POC study, denoted as a Phase IIa 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. After POC study, a DF study, denoted as a Phase IIb study, is often conducted to assess the efficacy with a relatively larger number of patients, and the primary endpoint is assessed at a later time point post-dosing. Similar to the situation as discussed by Chow et al (2007) and Chow and Tu (2008), the endpoints of the two studies are similar but different, e.g. a biomarker versus a regular clinical endpoint or the same study endpoint with different treatment durations. One opportunity of adaptation is to combine Phase IIa and Phase IIb, and another opportunity for adaptation is with DF studies. DF studies are usually conducted to find the optimal dose for phase III confirmatory trials, assessing both efficacy and safety. Conventionally, DF studies have a small, fixed number (3 or 4) of active doses, and try to choose the most promising dose among them. This approach works well with an important assumption that conventional studies don't always meet: the doses assessed include the optimal dose. The conventional approach also usually powers DF studies at the same level for every dosing arm, 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 arms (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 the optimal therapeutic level. This increases the risk of failure in the current phase or a more costly failure in a subsequent study, and poor 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). We suggest an adaptive design with seamless Phase IIa/IIb transition and enhanced dose finding approach.

In real situation at trial design stage, sample size or power is usually calculated based on 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 tolerance. With a couple of assumed utility models, the simulation results show that the proposed adaptive design tests more active doses with higher actual power and comparable or smaller sample size in a shorter overall duration, compared to a conventional design.

The rest of the paper is organized as follows. We introduce the concept of utility in Section 2, and elaborate the adaptive design with seamless Phase IIa/IIb transition and enhanced dose finding approach in Section 3. In Section 4, we show two hypothetical examples and compare the proposed adaptive design to the conventional design. We will conclude the paper in Section 5 with some discussion.

2. Utility

In normal situations, there is a threshold 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, see Figure 1. In addition to the desired drug effect, there is also side effect, which usually rises with the increase of dose levels. Therefore a higher dose with worse side effect can in general make it less desirable. We use utility to account for both therapeutic effect (efficacy) and side effect (safety and tolerance). 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 our simulation, we will use utility 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.

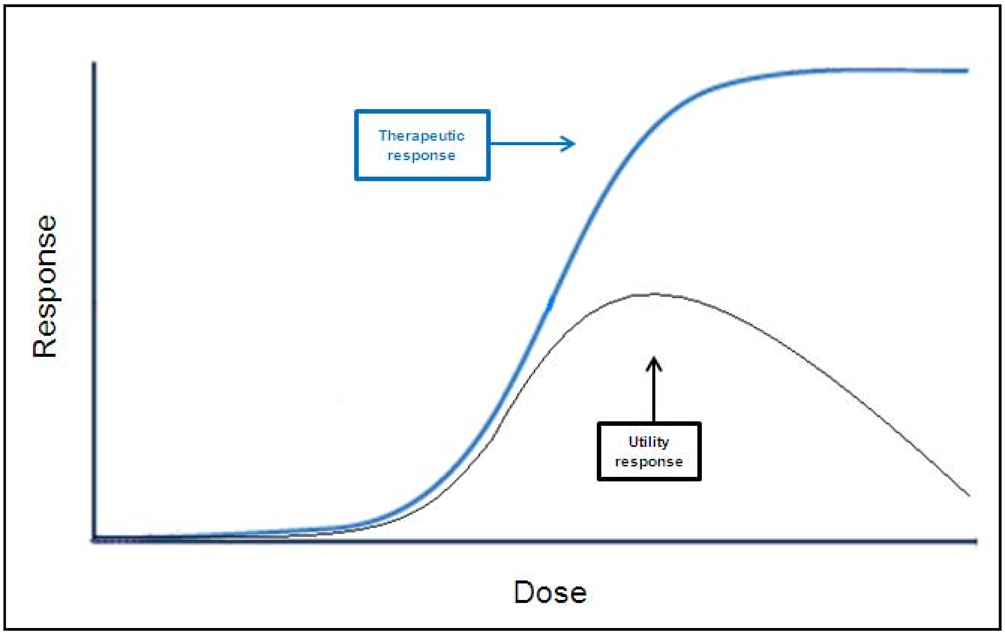

Figure 1. Dose response with therapeutic effect and utility

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

The endpoints of Phase IIa and IIb studies are often similar but different, e.g. a biomarker versus a regular clinical endpoint or the same study endpoint with different treatment durations. Considering an adaptive design that combines POC and DF, we denote the primary endpoint for the two stages with E_{poc} and E_{df}, respectively, which are correlated. In practice, oftentimes 2 active dose levels are kept in the phase III studies, and this is the scenario we will focus on.

To improve chances of correctly identifying optimal doses for phase III trials with economical testing of more doses in mind, we start with 6 parallel groups, i.e. 6 active dose levels, (the number of groups may increase if appropriate). Each 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.

Rationale for half patients in a delayed-treatment subgroup of every dose level:
Why not 6 active arms + 1 placebo arm in parallel with equal size? In our design, we conduct a POC test at an early stage, where we combine the active doses excluding a couple of 'losers', and compare to the placebo. To have a balance between the pooled active patients and placebo patients, we start with 6 groups, and each of them is balanced between active and placebo. After POC, only two delayed-treatment subgroups remain on placebo, and constitute the placebo group for the DF test later.

Stage 1: POC with 6 parallel groups (12 subgroups)

1) When 0.2×G patients in each subgroup are assessed for E_{poc}, find the worst 2 subgroups among the actively-treated, and the worst 2 subgroups among the placebo-treated. The worst performers are identified by comparing descriptive summary statistics, and no inferential analysis is performed. These worst performers will be excluded from the POC test.
2) With the better-performing 4 active subgroups (immediate-treatment subgroups) and 4 placebo subgroups (delayed-treatment subgroups), test the superiority of the pooled active doses over placebo, i.e., conduct the following test between the immediate-treatment patients pooled together (0.8×G) and the delayed-treatment patients pooled together (0.8×G) in the better-performing 4 active subgroups and 4 placebo subgroups. The significance level for this test is two-sided 0.05. Without loss of generality, we assume a greater value means better.

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

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 2 active doses are excluded. The worst 2 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.

2a) If superiority is not established, then POC fails, stop the study.
2b) Otherwise, drop the worst 2 active doses among the 6, and do the following to the delayed-treatment subgroups:

(i) **Convert to active**: for the 4 groups with better active doses, patients in the delayed-treatment subgroups convert to the corresponding active dose.

(ii) **Remain on placebo without interruption**: for the group with the worst performing active dose, patients in the delayed-treatment subgroup remain on placebo and continue subsequent assessments without interruption,

(iii) **Remain on placebo with dosing restart**: for the group with the second worst performing active dose, patients in the delayed-treatment subgroup also remain on placebo, but their dosing schedule will restart from the beginning.

The above (ii) and (iii) constitute the placebo arm for the final test. Therefore 5 arms remain for Stage 2.

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

Stage 2: DF with 5 arms

1) When 0.4×G patients in each of the remaining 4 active arms are assessed for E_{df}, drop the worst 2.

2) When G patients in each of the remaining 2 active arms and the placebo arm are assessed for E_{df}, conduct the following test to compare each active dose to placebo in the quantitative order, i.e. first test the dose with a better descriptive statistic. The significance level for the test of both doses is the same, two-sided 0.05.

$$H_{2n}: \mu_{a,df} - \mu_{p,df} \le 0 \text{ vs. } H_{2a}: \mu_{a,df} - \mu_{p,df} > 0 \qquad \text{(ii)}$$

2a) If superiority is established (for at least one active dose), the DF study succeeds and the best doses are found.
2b) If superiority cannot be established (i.e. null hypothesis cannot be rejected in either dose), the DF study fails.

With the suggested design, the sample size is 4.2×G, which is comparable to that of a conventional DF study with 3 active treatment arms and 1 placebo arm, 4×G. Figure 2 shows the study flowchart.

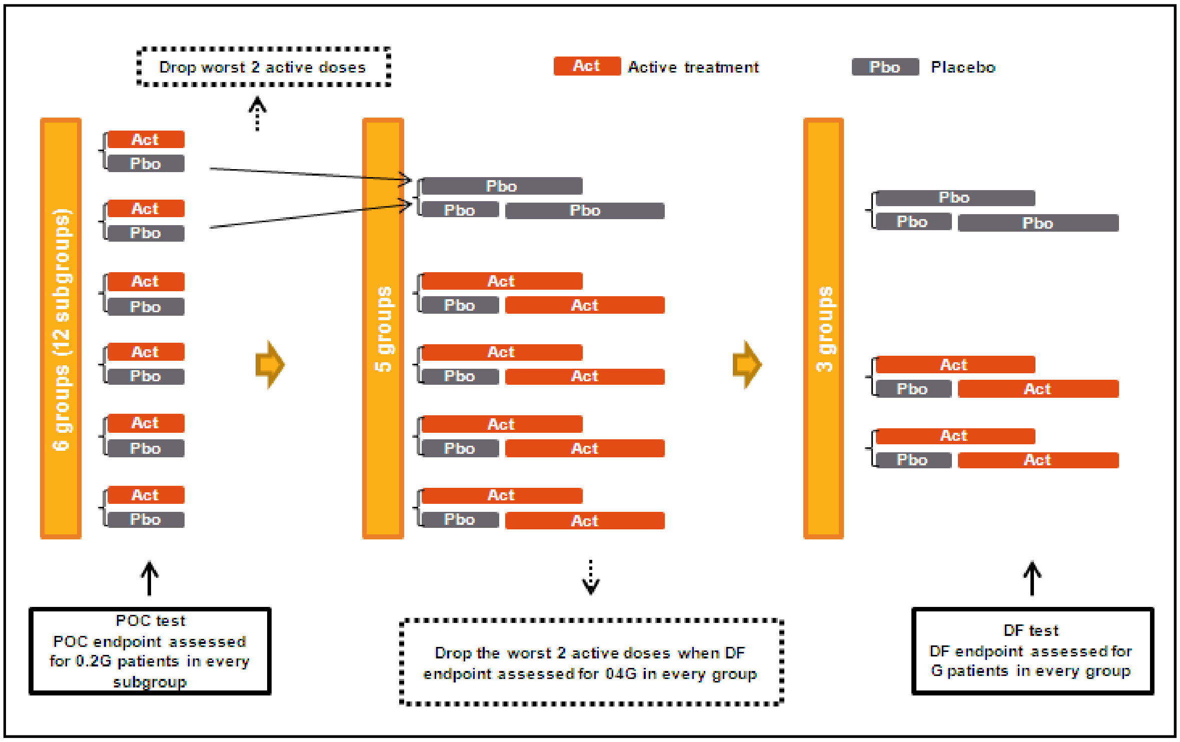
Drop worst 2 active doses
Act Active treatment
Pbo Placebo
6 groups (12 subgroups)
Act
Pbo
5 groups
Pbo
Pbo
Act
3 groups
POC test
POC endpoint assessed for 0.2G patients in every subgroup
Drop the worst 2 active doses when DF endpoint assessed for 04G in every group
DF test
DF endpoint assessed for G patients in every group

Figure 2. Study flowchart of the adaptive design

4. Hypothetical examples

In a usual case with the conventional approach, the POC study has 1 active arm of a high dose level, and the DF has 3 active arms. In the simulation, we consider a continuous primary endpoint, but the situation with a binary primary endpoint should be similar.

Two hypothesis tests are involved, one for POC and the other for DF. However, no multiplicity adjustment is made for each individual test. The results show that the type I error rate for the whole testing procedure is well controlled.

4.1. Good pick in conventional approach

In a conceptual fashion, let's assume the mean of the efficacy endpoint and the utility of the 6 active doses and the placebo is as shown in Figure 3. The common standard deviation is 1.20 for POC efficacy, 125 for POC utility, 1.76 for DF efficacy, and 137 for DF utility.

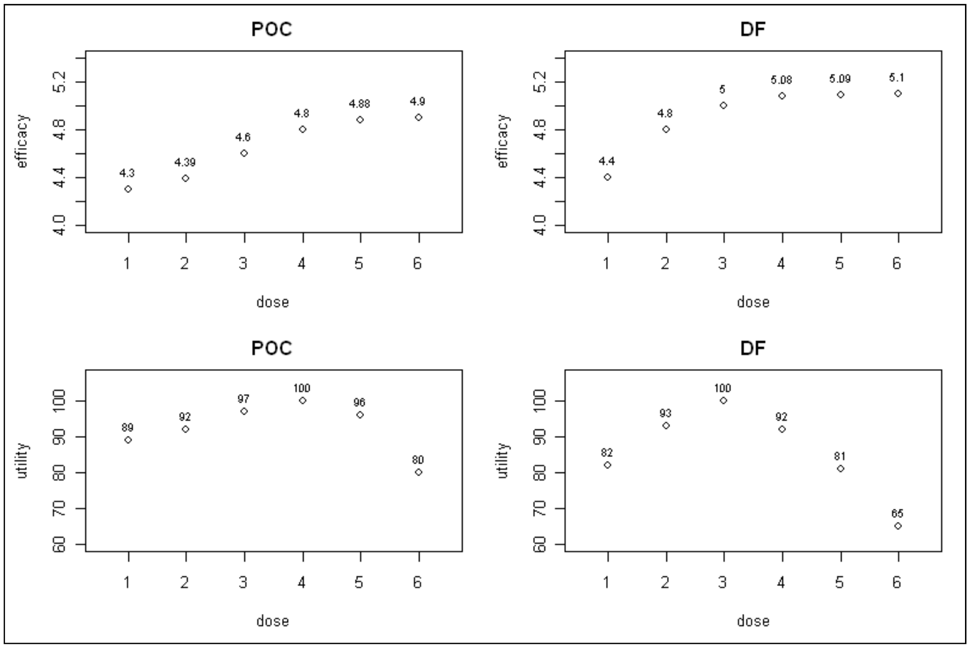

Figure 3. Efficacy and utility of the POC and DF endpoint in the good pick example

How the utility parameters are specified:
*The best dose (considering efficacy, safety and tolerance) has a utility of 100 for both POC and DF. For POC, as it is a short time point, safety and tolerance would not draw down much of the benefit, the worst dose still has a utility of 80. For DF, the worst dose has a lower utility of 65, because this is a longer time point where safety and tolerance issues can further undermine the therapeutic benefit. Placebo is assumed to have a utility of 20 at POC time point, and 15 at DF time point. The standard deviation is specified such that the actual power (power based on utility) with the best dose is 96% of the nominal power (i.e. 77%=0.80*0.96) for POC, and 92% of the nominal power (i.e. 74%=0.80*0.92) for DF.*

As an example of a good pick of doses with the conventional approach, assume Dose 4 is picked for POC study, and Doses 1, 3, 5 are picked for DF study. The sample size is calculated with the efficacy endpoint. With a target power of 80% for each study, the sample size is 74=2*37 for POC (active mean 4.8, placebo mean 4.0, common standard deviation 1.2, 1:1 randomization, alpha=0.025 one-sided, power=80%), and 200=4*50 for DF (active mean 5, placebo mean 4, common standard deviation 1.76, 1:1:1:1

randomization, alpha=0.025 one-sided, power=80%). The nominal power of success in both POC and DF studies is 64%=80%*80%.

For the conventional approach, the actual power, i.e. power in terms of utility, is 77% for POC study (active mean 100, placebo mean 20, common standard deviation 125, 1:1 randomization, alpha=0.025 one-sided, per arm sample size 37), and 74% for DF study (active mean 100, placebo mean 15, common standard deviation 137, 1:1:1:1 randomization, alpha=0.025 one-sided, per arm sample size 50). Therefore the actual power of the two studies together is 57%=77%*74%.

Now let's check the power in terms of utility for the proposed adaptive design. Assume the 6 active doses and the placebo have the same parameters as in Table 1. Different level of correlation between the POC and DF endpoints are checked. For each level of the correlation among 0, 0.5, 0.7, 0.9, and 1, conduct 10,000 runs of simulation, found the proportion that the POC is successful, i.e. POC power, and the proportion that both POC and DF are successful, i.e. Overall power. The results are in **Table 1**. The impact of correlation is negligible.

Correlation	POC power	Overall power
0	76.5%	71.4%
0.5	74.8%	70.6%
0.7	75.4%	71.3%
0.9	75.7%	71.6%
1	75.1%	70.9%

Table 1. POC power and Overall power for adaptive design with parameters in the situation of a good pick with conventional approach.

To check the type I error rate of the adaptive design in terms of utility, we assume the utility of each active dose is the same as placebo. For each level of the correlation, conduct 10,000 runs of simulation, and find the proportion that the POC is successful, i.e. POC type I error rate, and the proportion that both POC and DF are successful, i.e. Overall type I error rate. The results are in **Table 2**. The type I error rate is well controlled below required 0.025 level. Again, the impact of correlation is negligible.

Correlation	POC Type I error rate	Overall Type I error rate
0	0.0143	0.0010
0.5	0.0154	0.0013
0.7	0.0132	0.0010
0.9	0.0155	0.0013
1	0.0148	0.0013

Table 2. POC type I error rate and Overall type I error rate.

4.2. Bad pick with conventional approach

Now let's look at an example of a bad pick with the conventional approach. Assume the mean of the efficacy endpoint and the utility of the 6 active doses and the placebo is as shown in Figure 4. The common standard deviation is 1.20 for POC efficacy, 125 for POC utility, 1.76 for DF efficacy, and 137 for DF utility. Assume Dose 6 is picked for POC study, and Doses 1, 3, 5 are picked for DF study. As the sample size is calculated with efficacy endpoint, it will be the same as in the example of a good pick, i.e. 74=2*37 for POC and 200=4*50 for DF. But the actual power is much lower in this case with the conventional approach. Consider the utility parameters for these doses, the power is 53% for POC, and 65% for DF, and the Overall power of the two studies will be 34.4%=53%*65%.

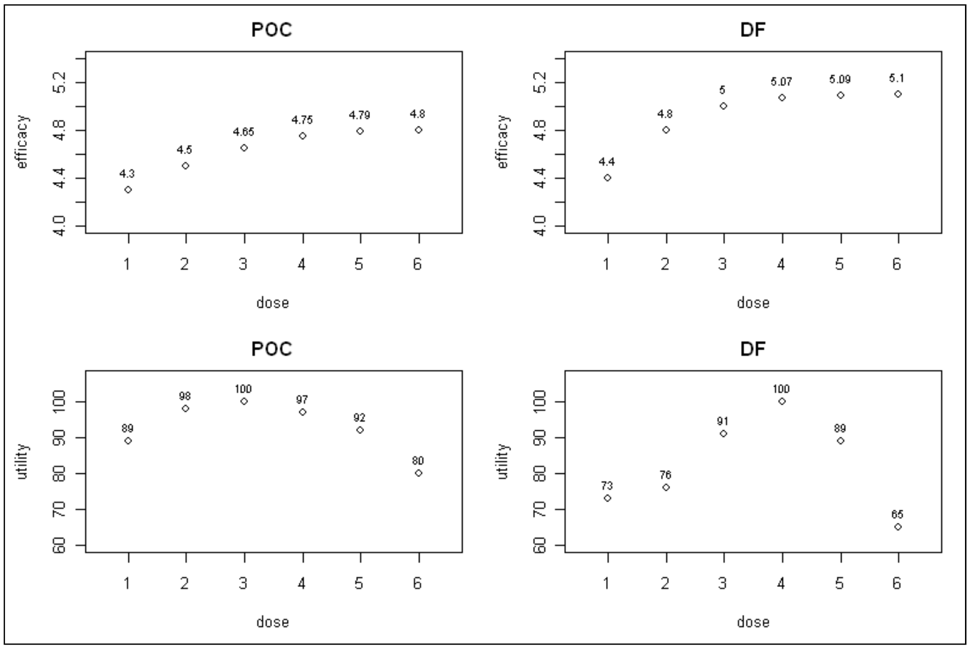

Figure 4. Efficacy and utility of the POC and DF endpoint in the bad-pick example

Similar to what have been done with the good pick example, the power results for the adaptive design are presented in **Table 3**. Again, the impact of correlation is slim.

Correlation	POC power	Overall power
0	77.9%	71.7%
0.5	77.6%	72.8%
0.7	77.0%	72.8%
0.9	78.1%	74.4%
1	78.6%	74.6%

Table 3. POC power and Overall power for adaptive design with parameters in the situation of a bad pick with conventional approach.

The type I error rate is similar to that in the good pick example, well controlled. (Data not shown.)

Table 4 shows the comparison between the adaptive design and the conventional design considering the examples of good pick and bad pick with the conventional approach. The results show that the proposed adaptive design tests more active doses with higher actual power and comparable or smaller sample size than required by a conventional design.

	Conventional design	Adaptive design
Actual POC power	57.0% ~ 77.0%	75.4% ~ 78.6%
Actual Overall power	34.4% ~ 57.0%	70.6% ~ 74.6%
Total number of patients	274	210

Number of active doses	4	6
Gap between POC and DF	Yes	No

Table 4: Comparison between adaptive design and conventional design**.**

5. Discussion

In the situation that only one active dose is maintained for the phase III trials, it will be the same in Stage 1 (for POC) as the scenario discussed in Section 2 where 2 active doses to remain in phase III, but it is a little different in Stage 2 (for DF).

1) When 0.4*G patients in each of the remaining 4 active arms are assessed for E_{df}, drop the worst among them.

2) When 0.6*G patients in each of the remaining 3 active arms are assessed for E_{df}, drop the worst among them.

3) When 0.8*G patients in each of the remaining 2 active arms are assessed for E_{df}, drop the worse of the two.

4) When G patients in the remaining 2 arms (the best active and the placebo) are assessed for E_{df}, test the superiority of the best active dose over the placebo, i.e., conduct the following test.

$$H_{20}: \mu_{b,df} - \mu_{p,df} \geq 0 \text{ vs. } H_{2a}: \mu_{b,df} - \mu_{p,df} < 0 \qquad \text{(ii)}$$

6a) If superiority is established, the DF study succeeds and the best dose is found.
6b) If superiority cannot be established, the DF study fails.

Acknowledgement

We thank Drs Peter Mesenbrink and Heinz Schmidli for their review and comments on an earlier version of this paper, which have been incorporated into this version. This work is partially supported through the National Institutes of Health (grant P01CA142538).

References

Chow, S. C., Lu Q., Tse S. K. (2007) Statistical analysis for two-stage seamless design with different study endpoints. J Biopharm Stat. 17(6):1163-76.

Chow, S. C., Tu Y. H. (2008) On Two-stage Seamless Adaptive Design in Clinical Trials. J Formos Med Assoc. 107(12 Suppl):52-60.

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