

Optimal decision criteria for the study design and sample size of a biomarker-driven phase III trial

Abstract

Background: The design and sample size of a phase III study for new medical technologies were historically determined within the framework of frequentist hypothesis testing.

Recently, drug development using predictive biomarkers, which can predict efficacy based on the status of biomarkers, has attracted attention, and various study designs using predictive biomarkers have been suggested. Additionally, when choosing a study design, considering economic factors, such as the risk of development, expected revenue, and cost, is important.

Methods: Here, we propose a method to determine the optimal phase III design and sample size and judge whether the phase III study will be conducted using the expected net present value (eNPV). The eNPV is defined using the probability of success of the study calculated based on historical data, the revenue that will be obtained after the success of the phase III study, and the cost of the study. Decision procedures of the optimal phase III design and sample size considering historical data obtained up to the start of the phase III study were considered using numerical examples.

Results: Based on the numerical examples, the optimal study design and sample size depend on the mean treatment effect in the biomarker-positive and -negative populations obtained from historical data, the between-trial variance of response, the prevalence of the biomarker-positive population, and the threshold value of probability of success required to go to phase III study.

Conclusions: Thus, the design and sample size of a biomarker-driven phase III study can be appropriately determined based on the eNPV.

Keywords :

predictive biomarkers, historical data, probability of success, net present value, go/no-go decision

Introduction

In recent years, the efficacy and safety requirements for new pharmaceutical products have increased; in this regard, a decrease in the success rate of drug development programs and the prolongation of the developmental period are regarded as problems. Drug development using biomarkers is considered one of the measures against these problems.¹ Predictive biomarkers are used to predict the difference in treatment efficacy based on the status of the respective biomarker. These predictive biomarkers can also identify the patient population with the highest efficacy, thus, reducing the required sample size of the study and increasing the possibility of detecting efficacy, compared to cases in which patient populations are not identified.^{2,3} Here, we focus on predictive biomarkers (hereinafter referred to as biomarkers) and propose a method for deciding the optimal study design and sample size for a phase III study using these biomarkers.

In general, the sample size in a phase III study is calculated based on the frequentist hypothesis testing. However, in terms of investment decisions on phase III studies, it is important to consider the economic factors important to the company, such as expected revenue, development costs, and examination periods, when determining the sample size.⁴

As the purpose of drug development is to deliver superior new drugs to patients more quickly, it is important to evaluate the probability of success (PoS) of the planned phase III study and use it in decision making. Kirchner et al.⁵ proposed a method for setting the optimal sample size for phase II studies and determine the go/no-go decision for the subsequent phase III study considering expected revenue, costs, and the PoS of phase II and the phase III studies.

Net present value (NPV) is defined as the present value of future profit minus the present value of investment cost. The NPV obtained by weighting the NPV of multiple scenarios considering the occurrence probability of each scenario is called the expected NPV (eNPV). These metrics are used for investment judgments. Patel et al.⁶ has proposed a method to maximize the value of the portfolio by determining the optimal sample size of a phase III study using the PoS and eNPV of the phase III study if the budget is limited.

When designing a phase III study using biomarkers, the efficacy in biomarker-positive and biomarker-negative populations obtained at the time of planning the study should be considered.³ Graf et al.⁷ and Ondra et al.⁸ defined utility functions taking future profit and costs of the phase III study into consideration from the perspectives of sponsors and public health in selecting study designs using biomarkers. Thus, they proposed a method to decide the optimal study design and sample size for a phase III study using utility.

Here, we propose a method to determine the optimal study design and sample size and judge whether a biomarker-driven phase III study should be conducted using the eNPV as an indicator. We expand on the methods of Graf et al.⁷ and Ondra et al.⁸ and propose to calculate the PoS, which is necessary for eNPV derivation, using historical data. Additionally, we propose to calculate the revenue required for eNPV by adapting the market share model of Kikuchi et al.⁹ and to use components of eNPV proposed by Patel et al.⁶ First, we describe a

motivating example and the considered phase III study designs. Next, we define eNPV and the proposed decision procedures to determine the optimal study design and sample size. Finally, we will determine the optimal phase III study design and sample size based on historical data obtained up to the start of the phase III study using numerical examples.

Motivating Example

Apolipoprotein E allele $\epsilon 4$ (ApoE4) has been reported as a risk factor for Alzheimer's disease.¹⁰ In a placebo-controlled phase II study of peroxisome proliferator-activated receptor gamma agonist, rosiglitazone, in patients with mild and moderate Alzheimer's disease, the results indicated that in the overall population, there was no improvement in the mean Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-cog), which was the primary endpoint, from baseline to 24 weeks. However, there was a significant improvement of ADAS-cog in the ApoE4-negative population and no improvement in the ApoE4-positive population. This result suggests that ApoE4 may be a biomarker for rosiglitazone efficacy in Alzheimer's disease.¹¹ However, in the clinical trials for acetylcholinesterase inhibitors, tacrine and donepezil in Alzheimer's disease, there were studies that showed efficacy only in the ApoE4-positive population or in both the ApoE4-positive and -negative populations.^{12,13,14} Therefore, uncertainty remains as to whether ApoE4 can predict differences in efficacy. In such a situation, it is important to appropriately set the design of the phase III study by considering information on the biomarkers and efficacy obtained, including that on other drugs used to treat the same disease.

Methods

Considered phase III trial designs

Here, we consider the realistic situation of planning a randomized phase III study. The purpose of this phase III study is to confirm that the experimental arm (E) has superior efficacy to the control arm (C). Within this study, only one binary biomarker that affects efficacy has been identified before the study. The full population (F) will be divided into the biomarker-positive population (P) and biomarker-negative population (N). The mean treatment effect of E against C in the population i ($i=F, P, N$) is denoted by Δ_i and σ^2 denotes the variance of an individual response in each arm and population, and is considered to be known ($\sigma = 1$). We assume that the estimates of the mean treatment effect $\hat{\Delta}_i$ are normally distributed. Note that throughout this study, $\Delta_P \geq \Delta_N$ and $\Delta_F = \lambda\Delta_P + (1 - \lambda)\Delta_N$ will be held as true for simplicity where λ is considered as the known prevalence of P among F .

Here, we consider four types of possible phase III study designs: traditional design, enriched design, stratified design, and adaptive enrichment design (Figure 1). The treatment effect in F and/or P are tested and the treatment effect in N is not tested in each study design. The outline of each study design is shown below.

[Insert Figure 1]

(1) Traditional design

In a traditional design, subjects are enrolled from F , and the hypothesis $H_{0,F}: \Delta_F \leq 0, H_{1,F}: \Delta_F > 0$ is tested. n denotes the sample size of F in each arm, and α denotes a one-sided

significant level ($\alpha=0.025$). The Z-statistic is given as $z_F = \hat{\Delta}_F \sqrt{n/2\sigma^2}$, and the hypothesis $H_{0,F}$ is rejected when $\hat{\Delta}_F \geq z_\alpha \sqrt{2\sigma^2/n}$; thus, the study is judged to be successful. Note that T_n refers to the traditional design with a sample size of n in each arm.

(2) Enriched design

In an enriched design, subjects are enrolled from P only, and the hypothesis $H_{0,P}: \Delta_P \leq 0, H_{1,P}: \Delta_P > 0$ is tested. n denotes the sample size of P in each arm. The Z-statistic is given as $z_P = \hat{\Delta}_P \sqrt{n/2\sigma^2}$, and the hypothesis $H_{0,F}$ is rejected when $\hat{\Delta}_P \geq z_\alpha \sqrt{2\sigma^2/n}$; thus, the study is judged to be successful. Note that E_n refers to the enriched design with a sample size of n in each arm.

(3) Stratified design

In a stratified design, subjects are enrolled from F and two hypotheses $H_{0,F}: \Delta_F \leq 0, H_{1,F}: \Delta_F > 0$ and $H_{0,P}: \Delta_P \leq 0, H_{1,P}: \Delta_P > 0$ are tested. Here, we adopt the closed testing procedure proposed by Spiessens and Debois¹⁵ for the multiple hypothesis testing method.

Let $\hat{\Delta}_P$ and $\hat{\Delta}_N$ be the estimates of mean treatment effect in P and N , respectively, and Δ_0 be the minimum clinically important mean treatment effect ($\Delta_0 = 0$). Here, we assume that a regulatory agency will require a certain efficacy for both P and N in order to approve F in the stratified design.^{8, 16} Therefore, the study is considered successful in F only when $H_{0,F}$ is rejected, and both $\hat{\Delta}_P \geq \Delta_0$ and $\hat{\Delta}_N \geq \Delta_0$ are satisfied.

We express the cases of success of the study in F and P as $S_{S,1}$ and $S_{S,2}$, respectively. Note that S_n refers to the stratified design with a sample size of n in each arm. Refer to the supplementary material for details on the considered stratified designs.

(4) Adaptive enrichment design

In an adaptive enrichment design, subjects are enrolled from F and an appropriate population group is selected at an interim analysis. Jenkins et al.¹⁷ proposed a method to select the population to be enrolled after the interim analysis and determine futility stopping by comparing the estimated value of efficacy in each population obtained in the interim analysis with a preset threshold value. Here, we refer to Jenkins et al.¹⁷ and consider the adaptive enrichment design, which determines the population selection and futility stopping at the interim analysis. To control the family-wise error rate, we adopt the combination test approach with the Spiessens and Debois test proposed by Freide et al.¹⁸

Let n be the sample size in each arm in the full population. An interim analysis will be conducted when the data of the subjects $n_1 = nt'$ ($0 < t' < 1$) in each arm are obtained ($t' = 0.5$). Let $\hat{\Delta}_{F,1}$ and $\hat{\Delta}_{P,1}$ be the estimates of the mean treatment effect in F and P , respectively, obtained based on the data of n_1 subjects. Let c_F and c_P be the threshold value of F and P , respectively ($c_F = c_P = 0.1$). Using these values, the decision rule at the interim analysis is as per Table 1.

Table 1. Decision rule at the interim analysis

[Insert Table 1]

In case 1 and 2, hypothesis testing is performed on F . However, similar to the stratified design, the study is judged to be successful in F when $H_{0,F}$ is rejected and both $\hat{\Delta}_P \geq \Delta_0$ and $\hat{\Delta}_N \geq \Delta_0$ are satisfied. We express the case of success of the study in F and P in case 1 as $S_{A1,1}$ and $S_{A2,1}$, respectively, and the case of success of the study in case 2 and case 3 as $S_{A,2}$ and $S_{A,3}$, respectively. Additionally, $S_{A,4}$ indicates the situation in which the study stopped for futility at the interim analysis. Note that A_n refers to the adaptive enrichment design with a sample size of n in each arm. Refer to the supplementary material for the details on the considered adaptive enrichment design.

Expected net present value

In general, NPV is defined as follows:

$$\text{NPV} = \text{discounted total revenue} - \text{discounted total cost}.$$

Marchenko et al.¹⁹ states that total revenue consists of the price of the drug, the population size, the market share, and years of revenue. Here, we define market size (M ; millions of US dollars (MUSD)/year) in F as the price of the drug $\times$ population size ($M=1000$). Note that sales and production costs are excluded from the price of the drug. We express the market size in P as λM . Market share (m ; %) is determined based on the efficacy obtained in the phase III trials. Kikuchi et al.⁹ proposed that market share be expressed as a logistic function with the posterior mean and posterior standard deviation of efficacy based on the results of the phase III trial as parameters. Here, we assume that market share depends on the mean treatment effect and its variance (a and b , respectively) when the phase III study is successful. Market share is defined by the following formula

$$m(a, b) = \frac{1}{1 + \emptyset \exp \left\{ \frac{2\beta(1.5b - a)}{B - A} \right\}}, \emptyset = \exp \left\{ \frac{\beta(A + B)}{B - A} \right\},$$

where A is the mean treatment effect that some patients consider switching to new drugs with a probability of 87% or more and B is the mean treatment effect that all patients consider switching to a new drug with a probability of 87% or more. We set $A=0.1$ and $B=0.5$. β is the slope of the logistic function, and $\beta=1$. In practice, parameters A and B are unknown; therefore, it is necessary to investigate the preference of treatment selection with some key opinion leaders and determine its value. Note that Kikuchi et al.⁹ set the multiplier 1.5 in $m(a,b)$ in their paper, but it is an arbitrary value. If we want to change the probability from 87 % to 95 %, we need to change the multiplier from 1.5 to 1.96. For simplicity, we assume that market share is constant regardless of the period. Additionally, the years of revenue is defined as the remaining protected period of the patent at the start of the phase III study minus the period from the start of the phase III study to the launch of the new drug.

Here, the total cost (MSUD) is the cost required to conduct a phase III study. We define eNPV below with a consideration that the total revenue, which will only be obtained after a successful phase III study. The annual discount rate used to calculate the discounted total revenue and discounted total cost is denoted by ρ , and $\rho=0.1$.²⁰

$$\begin{aligned} \text{eNPV (MUSD)} &= PoS \times \text{market size} \times \text{market share} \times \text{discounted years of revenue} \\ &\quad (\text{discounted total revenue}) - \text{discounted phase III cost.} \end{aligned}$$

Let R_{d_n} ($d = T, E, S, A$) and C_{d_n} be the expected discounted total revenue and discounted phase III cost, respectively, for each design with sample size n per arm. The eNPV in each design will be given by the following formula:

$$eNPV_{d_n} = R_{d_n} - C_{d_n}.$$

Calculation of probability of success based on historical data

Chuagn-Stein²¹ proposed PoS as a weighted mean of power in the possible range of treatment effect (Δ), and formulated PoS as follows:

$$PoS = \int P(success|\Delta)p(\Delta|Historical\ Data)d\Delta;$$

where, $P(success|\Delta)$ is the probability of success of the study based on the treatment effect, and $p(\Delta|Historical\ Data)$ is the posterior distribution of Δ based on the results of the previous study. Chuagn-Stein²¹ only used the result of the previous study to derive the distribution of Δ , but when available historical data exist, it becomes possible to derive a more accurate distribution of Δ by integrating this information. Note that the PoS can be calculated as the prior expectation of the power and O'Hagan et al.²² defined and named this as an assurance.

A method to derive the predictive distribution of efficacy of the next phase study based on historical data was proposed by Neuenschwander et al.²³ Here, we apply the method presented by Neuenschwander et al.²³ to the mean treatment effect obtained as historical data and define the predictive distribution of the mean treatment effect in this phase III study.

Now, assume that there are H historical trials available. Let Δ_i^h be the parameter of the mean treatment effect in population i of historical trial h ($h=1, \dots, H$), while $\hat{\Delta}_i^h$ is its estimated value.

Let Δ_i^* be the parameter of the mean treatment effect in population i of the phase III study, while $\hat{\Delta}_i^*$ is its estimated value. Using these parameters, we assume the following model: n_i^h indicates the sample size of each arm in population i of historical trial h , and τ indicates the between-trial standard deviation of response in each arm. Note that τ is known for simplicity.

$$\hat{\Delta}_i^h | \Delta_i^h \sim N(\Delta_i^h, 2\sigma^2/n_i^h),$$

$$\Delta_i^1, \dots, \Delta_i^H, \Delta_i^* | \mu, \tau^2 \sim N(\mu, 2\tau^2).$$

Assuming a uniform distribution in the hyperprior distribution of μ , the following equation is given:

$$\Delta_i^* | \Delta_i^1, \dots, \Delta_i^H, \tau^2 \sim N\left(\frac{\sum w_i^h \hat{\Delta}_i^h}{\sum w_i^h}, \frac{1}{\sum w_i^h} + 2\tau^2\right), w_i^h = \frac{1}{2\sigma^2/n_i^h + 2\tau^2}.$$

From above, the predictive distribution of treatment effect in population i (Δ_i^*) is given by the following equation:

$$\begin{aligned} \Delta_P^* &\sim N\left(\frac{\sum w_P^h \hat{\Delta}_P^h}{\sum w_P^h}, \frac{1}{\sum w_P^h} + 2\tau^2\right), & \Delta_N^* &\sim N\left(\frac{\sum w_N^h \hat{\Delta}_N^h}{\sum w_N^h}, \frac{1}{\sum w_N^h} + 2\tau^2\right), \\ \Delta_F^* &\sim N(\lambda\Delta_P^* + (1-\lambda)\Delta_N^*, \lambda^2 Var(\Delta_P^*) + (1-\lambda)^2 Var(\Delta_N^*)). \end{aligned} \quad (1)$$

The PoS of each design in sample size n in each arm is expressed by Pos_{d_n} . See Appendix 1 for the PoS for each design.

For simplicity, we assume that τ is known. However, τ will not be known in general, especially in the case where there is a limited amount of historical data. In such a case, if the experimental drug has other indications, those data may support the selection of τ . Additionally, it is important to conduct the sensitivity assessment using different priors for τ and to check how the conclusions change²³. As a solution to the selection of τ , calculation of the POS from historical data without using τ can be considered. Ibrahim et al.²⁴ proposed a method to calculate the POS, incorporating historical data using the power prior. Their method

can be used in the situations which has one or two historical data. Using their method to calculate the POS will be one of the solutions.

Calculation of expected discounted total revenue

The revenue is obtained only when the phase III study is successful, that is, statistical significance is verified. We assume that the market share depends on the mean treatment effect and its variance when the phase III study is successful. Therefore, we need to calculate these values to calculate the expected total revenue.

In the case of success in a traditional design, the expected value (E_T) and variance (V_T) of the mean treatment effect are calculated by the expected value and variance of a left-truncated normal distribution, which is left truncated in $z_{\alpha/2}\sqrt{2\sigma^2/n}$.²⁵ The period from planning the study to first patient in is expressed as s (month); the recruitment speed for F is expressed as r (subjects/month); the subject's follow up time is expressed as f (month), and the examination period, which includes the period until submission, is expressed as e (month). We set $s=3$, $r=40$, $f=6$, and $e=18$. Based on these variables, the period from planning the phase III study to launch (T'_T) is given by $T'_T = s + \frac{2n}{r} + f + e$. When T is the remaining period of the patent life at the start of the phase III study, the expected discounted total revenue (R_{T_n}) is given by the following formula:^{6,26}

$$R_{T_n} = PoS_{T_n} \cdot M \cdot m(E_{T_n}, V_{T_n}) \cdot \int_{T'_T=s+\frac{2n}{r}+f+e}^T e^{-\frac{\rho}{12}t} dt.$$

We set $T=120$. See Appendix 2 for the expected discounted total revenue for the other designs.

Calculation of discounted phase III cost

The costs (MUSD) associated with conducting a phase III study are the subject cost (c_{sub}), trial cost incurred each month during the study (c_{trial}), biomarker examination cost per subject (c_{bio}), companion diagnostics development costs (c_{CDx}), and the interim analysis implementation cost ($c_{interim}$). We set $c_{sub}=0.08$, $c_{trial}=3$, $c_{bio}=0.001$, $c_{CDx}=10$ and $c_{interim}=1.5$. We assume that c_{sub} is incurred during the months in which each subject is enrolled and c_{trial} is incurred every month beginning from month that the first patient is enrolled to the end of last patient's follow up. Next, we assume that c_{CDx} is incurred at the start of the phase III study and c_{bio} is incurred during the months in which each subject is enrolled in the enriched design, stratified design, or adaptive enrichment design studies, all of which all require a biomarker test. Finally, we assume that $c_{interim}$ is incurred when all subjects required for the interim analysis have completed the follow-up in an adaptive enrichment design that performs interim analysis. Based on the aforementioned variables, the discounted phase III cost, C_{T_n} , in a traditional design is given by the following formula:^{6,26}

$$C_{T_n} = rc_{sub} \int_s^{s+\frac{2n}{r}} e^{-\frac{\rho}{12}t} dt + c_{trial} \int_s^{s+\frac{2n}{r}+f} e^{-\frac{\rho}{12}t} dt.$$

See Appendix 3 for the discounted phase III cost for the other designs.

Selecting the optimal design and sample size

From the perspective of the sponsor, eNPV and PoS are important when deciding whether to conduct the phase III study. It would be natural for the sponsor to secure the PoS above a certain value (c_{PoS}). However, if the PoS cannot be secured, the risk is considered high, and

the drug development is discontinued. If $eNPV_{d,max}$ denotes the maximum eNPV in the design d under the condition that $Pos_{d_n} \geq c_{Pos}$, the optimal design d^{opt} is given by:

$$d^{opt} = \operatorname{argmax}_{d \in D} \{eNPV_{d,max}\}.$$

The sample size determined based on $eNPV_{d,max}$ is defined as the optimal sample size. Note that when the $eNPV_{d,max}$ of d^{opt} is less than 0, the phase III study is not conducted, and drug development is discontinued.

Numerical examples

We consider a randomized parallel-group phase III study to confirm the efficacy of the experimental drug against the control drug. The optimal design and the sample size of the phase III study are strongly influenced by the efficacy of P and N in the historical data; thus, several scenarios are considered as historical data. To calculate the eNPV, it is necessary to set the parameters related to the study design, development cost, and revenue. Thus, we change the values of the parameters considered to affect the decision of the optimal design and sample size and fix other parameters to the appropriate values. The optimal design and sample size in each scenario were examined by 30,000 Monte Carlo simulations.

Scenario settings

We assume four scenarios (Table 2) as efficacy in historical data.

Table 2. Scenarios of efficacy in historical data

[Insert Table 2]

In scenario 1, the treatment is efficacious in P and N . In scenario 2, the efficacy in N is half of that in P . In scenario 3, the efficacy in N is same as that in the control arm. In scenario 4, the efficacy in N is inferior to that of the control arm. In the derivation of the predictive distributions mentioned above, we assume a situation in which multiple historical data are available. However, to consider the impact of the magnitude of efficacy in historical data on the study design, we assume a situation in which the historical data only includes information of the phase II study using the same drug as that in the phase III study. It is assumed that the prevalence P (λ) is the same between the phase II and phase III studies.

Parameter settings

The sample size of the phase III study is calculated based on the magnitude of the expected efficacy, the variation of the efficacy, and the prevalence of P . The prevalence of P was tested at $\lambda=0.2, 0.5$, and 0.8 , whereas the between-trial standard deviation was tested at $\tau=0.05$ and 0.2 . Neuenschwander B et al.³² defined a small between-trial variance of response as $\sigma/\tau=20$, and a large between-trial variance of response as $\sigma/\tau=5$ (i.e. when $\sigma=1$, then $\tau=0.05$ and $\tau=0.2$, respectively). Additionally, because the PoS threshold value required to go to phase III is an important evaluation criteria for proceeding to the phase III study, c_{PoS} was tested as $c_{PoS}=0.7$ and 0.8 . The other parameters were set at fixed values as mentioned above. Based on the values of the fixed parameters, the maximum recruitment period and the maximum sample size in the two arms were calculated as 93 months and 3720, respectively. The sample size in the two arms varied by 10, between 10 and 3720. For a better understanding of the simulation settings, each parameter value used in the simulation is listed in Table A1 of Appendix 4.

Results

The maximum eNPV and optimal sample size of each design from scenario 1 to scenario 4 are shown in Figure 2a to Figure 2d, respectively. We show how the optimal design varies depending on the between-trial standard deviation (τ), prevalence of P (λ), and the PoS required to go to phase III (c_{PoS}) in each scenario. For a better understanding of the results, figures of PoS and sample size in each design and in each scenario are presented in the supplementary material.

[Insert Figure 2a]

[Insert Figure 2b]

[Insert Figure 2c]

[Insert Figure 2d]

Scenario 1 : $(\hat{\Delta}_P^1, \hat{\Delta}_N^1) = (0.4, 0.4)$

Regardless of the values of c_{PoS} , τ , and λ , the maximum eNPV of the traditional design was the highest and was selected as the optimal design. In scenario 1, the treatment is equally efficacious in P and N . Accordingly, the PoS in the maximum eNPV of the traditional design, which tests the full population, is equal to or higher than that of other designs, and the optimal sample size is the smallest. The traditional design was selected as the optimal design because the period to obtain revenue is long due to the small optimal sample size and the cost is low due to the short study period.

Scenario 2 : $(\hat{\Delta}_P^1, \hat{\Delta}_N^1) = (0.4, 0.2)$

In scenario 2, the efficacy for the treatment in N is half that in P . As a result, the stratified and adaptive enrichment designs were selected as the optimal designs. These two designs have higher PoS compared to that of the traditional design and the enriched design under the same sample size. The magnitude of the between-trial standard deviation affects the selection, of which the stratified design or the adaptive enrichment design is the optimal design.

In the case of $\tau=0.05$ and $\lambda \geq 0.5$, regardless of c_{PoS} , the maximum eNPV of the stratified design was slightly higher than those of the traditional and adaptive enrichment designs; thus, the stratified design was selected as the optimal design. If the efficacy is different between P and N , the sample size needed to increase the PoS in the stratified design and the adaptive enrichment designs, which is the sum of PoS of F and P , is smaller than that of the traditional design. Additionally, when the treatment is efficacious in P and N and the between-trial variance of response is small, the PoS of the stratified design will be slightly higher than that of the adaptive enrichment design. When the between-trial variance of the response is small, it is assumed that it is effective in F and/or P in scenario 2. Therefore, the stratified design that did not conduct the interim analysis was selected as the optimal design. Contrastingly, when $\tau=0.2$, the adaptive enrichment design was selected as the optimal design when $c_{PoS} = 0.7$. In the adaptive enrichment design, the PoS is often higher than that of the stratified design. When the between-trial variance of the response is large, the variances of predictive distribution of the treatment effect in P and N are large. Therefore, the adaptive enrichment design, which selects a more appropriate population at the interim analysis, considering historical data and information obtained up to the interim analysis, had higher PoS than the stratified design and was selected as the optimal design. If $c_{PoS} = 0.8$, no design in this scenario will be able to

secure this threshold, even with a large sample size. Thus, the phase III study will not be conducted.

Scenario 3 : $(\hat{\Delta}_P^1, \hat{\Delta}_N^1) = (0.4, 0.0)$

In scenario 3, the efficacy for the treatment in N is the same as that in the control arm. As a result, the stratified and adaptive enrichment designs were selected as the optimal designs. The main reason for this result is the same as that described in scenario 2.

First, we focus on a case in which $\tau=0.05$ and $\lambda \leq 0.5$. The efficacy in F is predicted to be less than half of the efficacy in P . Based on these variables, the PoS in the maximum eNPV of the adaptive enrichment design is greater than or equal to that of the stratified design. Additionally, the optimal sample size of the adaptive enrichment design is less than or equal to that of the stratified design. Therefore, the adaptive enrichment design was selected as the optimal design. However, when $\lambda=0.8$, the stratified design was selected as the optimal design, because high efficacy is expected even in F .

Next, we focus on a case in which $\tau=0.2$ and $c_{PoS}=0.7$. In this case, as in scenario 2, the adaptive enrichment design is selected as the optimal design regardless of λ . If $c_{PoS}=0.8$, the phase III study will not be conducted.

Scenario 4 : $(\hat{\Delta}_P^1, \hat{\Delta}_N^1) = (0.4, -0.2)$

In scenario 4, the efficacy in N is inferior to that of the control arm. Intuitively, one would think that the enriched design would be selected as the optimal design. However, in the cases where $\lambda \geq 0.5$, the adaptive enrichment design was selected as the optimal design. This is

because, where $\lambda \geq 0.5$, the predictive mean treatment effect in F exceeds 0.1 and there is a possibility that the observed treatment effect in N is positive owing to uncertainty of the treatment effect. The PoS in the maximum eNPV of the adaptive and enriched designs were the same and the optimal sample size in P was almost the same. Therefore, the adaptive enrichment design, which has a probability of getting approval in F, had a higher eNPV than the enriched design and was selected as the optimal design.

Note that the maximum eNPV of the adaptive enrichment design was greater than or equal to that of the stratified design, regardless of λ . This is because the PoS in the maximum eNPV of the adaptive enrichment design is greater than or equal to that of the stratified design and the optimal sample size of the adaptive enrichment design is less than or equal to the optimal sample size of the stratified design, regardless of λ .

In scenario 4, the inclusion of subjects in N into the phase III study may raise ethical arguments. However, uncertainties remain in the results obtained up to the phase III study, and therefore, confirmation of treatment effect is performed in the phase III study. In this case, it will be important to adopt the adaptive enrichment design and assess the treatment effect at the interim analysis and select the optimal population that should be tested. Consultation with a regulatory agency will be important before starting the phase III study.

Application to the motivating example

Here, we show what decisions are made when we apply the proposed method to the motivating example described above. Table 3 shows the result of clinical trial that evaluated the change from baseline in ADAS-cog compared to placebo in a randomized study. Note that a negative number indicates the improvement from the observed ADAS-cog at baseline. The

estimates of variance of treatment effect in each ApoE4 status were calculated using the sample size in each arm and the result to derive a predictive distribution of treatment effect.

Table 3. Result of clinical trial in the motivating example

[Insert Table 3]

Using the value presented in Table 3, we calculated the standard deviation of response in each arm and each population as $\sigma=4.94$ and we set the between-trial standard deviation as $\tau=0.25$, a small between-trial variance of response ($\sigma/\tau=20$) , and $\tau=0.99$, a large between-trial variance ($\sigma/\tau=5$). The predictive distributions of treatment effects in ApoE4-positive, ApoE4-negative, and full populations are derived by using formula (1) and above values, and presented in Table 4. Note that the prevalence of the ApoE4-positive population λ and ApoE4-negative population $(1-\lambda)$ are set as 0.65 and 0.35, respectively.²³

Table 4. Predictive distribution of treatment effects in each population

[Insert Table 4]

To calculate the eNPV, we also need to set the parameters related to the study design, the development cost, and the revenue. The values of each parameter used to calculate the eNPV are presented in Table A3 of Appendix 4. Table 5 shows the optimal sample size, PoS, and eNPV in each design.

Table 5. The optimal sample size, PoS, and eNPV in the motivating example

[Insert Table 5]

The eNPV was higher in the order of the adaptive enrichment design, stratified design, enriched design, and traditional design, regardless of the value of τ . Therefore, the adaptive enrichment design is the optimal design in the motivating example and we decided to use the adaptive enrichment design in the phase III confirmatory study.

Discussion

Based on the simulation results shown in the numerical examples, the optimal study design and sample size depend on the difference of efficacy in P and N (obtained based on historical data), between-trial variance, the prevalence of P , and the threshold value of PoS required to go to phase III study. Therefore, when applying the proposed method, it is important to conduct a sensitivity analysis on these parameters to evaluate how the optimal test design and sample size change. Additionally, among the study designs that were not selected as the optimal design, there was a design that was considered appropriate to be an optimal design considering the ease of implementation of the study in which maximum eNPV was not much different from the maximum eNPV of the optimal design. However, the proposed method does not consider the ease of design implementation when deciding an optimal study design. Therefore, a method that considers the ease of design implementation could be developed. Parameters specifying the allowable range of maximum eNPV differences between study designs (e.g. study design in which the maximum eNPV is larger than $0.95 \times$ the maximum eNPV of the optimal design) and, thus, specifying the study design that is the best and yet most simple to

implement could be determined. The method of deciding the optimal study design considering the ease of conducting the study in addition to the eNPV is a future research topic.

The choice of threshold value of PoS required to go to Phase III is important and may need discussion. From an investment perspective, if eNPV is the highest even when the PoS is small, it is possible to invest. For example, if an experimental drug is expected to be highly safe compared to the standard of care and there is a great potential benefit in case of success, a small PoS will be acceptable.

Here, we proposed to use the hierarchical model to calculate the PoS using historical data. Using this model, we will obtain one predictive mean treatment effect. However, there remains uncertainty in the treatment effect. In this case, modeling the uncertainty of the treatment effect through a Bayesian prior can be considered. Ondra et al.⁸ set several scenarios for treatment effect in P and N, and set a prior probability to each scenario. Our proposed method can be further expanded by applying their ideas. For example, assign the highest prior probability to the scenario which is derived from the historical data. The eNPV used for decision-making can be obtained by calculating the eNPV for each scenario and summing the eNPV weighted by the prior probability of each scenario.

Here, we did not consider safety information for the calculation of expected revenue, which is one of the components of the eNPV. For some drugs, the market share depends on safety as well as efficacy. Therefore, it is necessary to model market share considering safety. Kikuchi and Gittins²⁸ proposed a method to model market share with efficacy and safety results obtained in the phase III study. It would be possible to extend the method proposed in this paper by applying their method.

Finally, from a regulatory perspective, we propose that it will be important for them to consider an accuracy of treatment effect and perform a benefit and risk evaluation, including safety as well as efficacy in P and N, when reviewing the results of clinical trials using the predictive biomarker.

Conclusion

Thus, we have shown that the design and sample size of a phase III study in drug development using predictive biomarkers can be properly decided based on the eNPV.

Reference

1. FDA. Critical Path Opportunities List. 2006. <http://wayback.archive-it.org/7993/20180125035414/https://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/ucm076689.htm>.
2. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical Trial Designs for Predictive Marker Validation in Cancer Treatment Trials. *J Clin Oncol*. 2005;23:2020-2027.
3. FDA. Enrichment Strategies for Clinical Trials to Support Determination of Effectiveness of Human Drugs and Biological Products. 2019.
4. Patel NR, Ankolekar S. A Bayesian approach for incorporating economic factors in sample size design for clinical trials of individual drugs and portfolios of drugs. *Stat Med*. 2007;26:4976-4988.
5. Kirchner M, Kieser M, Götte H, Schüller A. Utility-based optimization of phase II/III programs. *Stat Med*. 2016;35:305-316.

6. Patel NR, Ankolekar S, Antonijevic Z, Rajicic N. A mathematical model for maximizing the value of phase 3 drug development portfolios incorporating budget constraints and risk. *Stat Med*. 2013;32:1763-1777.
7. Graf AC, Posch M, Koenig F. Adaptive designs for subpopulation analysis optimizing utility functions. *Biom J*. 2015;57:76-89.
8. Ondra T, Jobjörnsson S, Beckman RA, et al. Optimizing Trial Designs for Targeted Therapies. *PLoS One*. 2016;11:e0163726. doi:10.1371/journal.pone.0163726.
9. Kikuchi T, Pezeshk H, Gittins J. A Bayesian cost-benefit approach to the determination of sample size in clinical trials. *Stat Med*. 2008;27:68-82.
10. Tsai MS, Tangalos EG, Petersen RC, et al. Apolipoprotein E: Risk Factor for Alzheimer Disease. *Am J Hum Genet*. 1994;54:643-649.
11. Risner ME, Saunders AM, Altman JFB, et al. Efficacy of rosiglitazone in a genetically defined population with mild-to-moderate Alzheimer's disease. *Pharmacogenomics J*. 2006;6:246-254.
12. Poirier J, Delisle MC, Quirion R, et al. Apolipoprotein E4 allele as a predictor of cholinergic deficits and treatment outcome in Alzheimer disease. *Proc Natl Acad Sci USA*. 1995;92:12260-12264.
13. Greenberg SM, Tennis MK, Brown LB, et al. Donepezil therapy in clinical practice: A Randomized Crossover Study. *Arch Neurol*. 2000;57:94-99.
14. Winblad B, Engedal K, Soininen H, et al. A 1-year, randomized, placebo-controlled study of donepezil in patients with mild to moderate AD. *Neurology*. 2001;57:489-495.
15. Spiessens B, Debois M. Adjusted significance levels for subgroup analyses in clinical trials. *Contemp Clin Trials*. 2010;31:647-656.

16. Millen BA, Dmtirienko A, Ruberg S, Shen L. A Statistical Framework for Decision Making in Confirmatory Multipopulation Tailoring Clinical Trials. *Drug Inf J*. 2012;46:647-656.
17. Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharm Stat*. 2011;10:347-356.
18. Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Stat Med*. 2012;31:4309-4320.
19. Marchenko O, Miller J, Parke T, Perevozskaya I, Qian J, Wang Y. Improving Oncology Clinical Programs by Use of Innovative Designs and Comparing Them via Simulations. *Ther Innov Regul Sci*. 2013;47:602-612.
20. Baras AI, Baras AS, Schulman KA. Drug development risk and the cost of capital. *Nat Rev Drug Disc*. 2012; 11: 347-348
21. Chuang-Stein C. Sample size and the probability of a successful trial. *Pharm Stat*. 2006;5:305-309.
22. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharm Stat*. 2005;4:187-201
23. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. *Clin Trials*. 2010;7:5-18.
24. Ibrahim JG, Chen MH, Lakshminarayanan M, Liu GF, Heyse JF. Bayesian probability of success for clinical trials using historical data. *Stat Med*. 2015;34:249-264.
25. Jawitz JW. Moments of truncated continuous univariate distributions. *Adv Water Resour*. 2004;27:269-281.

26. Buser SA. LaPlace Transforms as Present Value Rules: A Note. *J Finance*. 1986;41:243-247.
27. Kikuchi T, Gittins J. A behavioral Bayes method to determine the sample size of a clinical trial considering efficacy and safety. *Stat Med*. 2009;28:2293-2306.
28. Farrer LA, Cupples LA, Haines JL, et al. Effects of Age, Sex, and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease. A Meta-analysis. *JAMA*. 1997;278:1349-56.
29. Ranina P, Santaguida P, Ismaila A, et al. Effectiveness of cholinesterase inhibitors and memantine for treating dementia : evidence review for a clinical practice guideline. *Ann Intern Med*. 2008;148:379-97.

Appendix 1. Formulation of PoS

The PoS of the traditional and enriched designs is formulated as follows, using the predicted distribution of treatment effect in each population of equation (1).

$$\begin{aligned}
 PoS_{T_n} &= \int_{-\infty}^{+\infty} P\left(\hat{\Delta}_F \geq z_\alpha \sqrt{\frac{2\sigma^2}{n}} |\Delta_F^*\right) f(\Delta_F^* | \Delta_P^1, \dots, \Delta_P^H, \Delta_N^1, \dots, \Delta_N^H, \tau^2) d\Delta_F^* \\
 &= 1 - \Phi \left[\frac{z_\alpha \sqrt{\frac{2\sigma^2}{n}} - \left(\lambda \frac{\sum w_{P,i} \hat{\Delta}_{P,i}}{\sum w_{P,i}} + (1 - \lambda) \frac{\sum w_{N,i} \hat{\Delta}_{N,i}}{\sum w_{N,i}} \right)}{\sqrt{\lambda^2 \left(\frac{1}{\sum w_{P,i}} + 2\tau^2 \right) + (1 - \lambda)^2 \left(\frac{1}{\sum w_{N,i}} + 2\tau^2 \right)}} \right], \\
 PoS_{E_n} &= \int_{-\infty}^{+\infty} P\left(\hat{\Delta}_P \geq z_\alpha \sqrt{\frac{2\sigma^2}{n}} |\Delta_P^*\right) f(\Delta_P^* | \Delta_P^1, \dots, \Delta_P^H, \tau^2) d\Delta_P^*
 \end{aligned}$$

$$= 1 - \Phi \left[\frac{z_\alpha \sqrt{\frac{2\sigma^2}{n}} - \frac{\sum w_{P,i} \hat{\Delta}_{P,i}}{\sum w_{P,i}}}{\sqrt{\frac{1}{\sum w_{P,i}} + 2\tau^2}} \right].$$

The PoS of the stratified and adaptive enrichment designs is calculated by simulations. The PoS in these study designs is represented as follows based on the study success criteria in each design mentioned above.

$$\begin{aligned} PoS_{S_n} &= P(S_{S,1} | \Delta_P^*, \Delta_N^*) (= PoS_{S_{n,1}}) + P(S_{S,2} \text{ only} | \Delta_P^*) (= PoS_{S_{n,2}}), \\ PoS_{A_n} &= P(S_{A1,1} | \Delta_P^*, \Delta_N^*) (= PoS_{A1_{n,1}}) + P(S_{A2,1} \text{ only} | \Delta_P^*) (= PoS_{A2_{n,1}}) \\ &\quad + P(S_{A,2} | \Delta_P^*, \Delta_N^*) (= PoS_{A_{n,2}}) + P(S_{A,3} | \Delta_P^*) (= PoS_{A_{n,3}}). \end{aligned}$$

Appendix 2. Expected discounted total revenue

In the case of success in the enriched design, the expected value (E_E) and variance (V_E) of efficacy are calculated in the same manner as that in the traditional design. The expected discounted total revenue of enriched design (R_{E_n}) is given by:

$$R_{E_n} = PoS_{E_n} \cdot \lambda M \cdot m(E_{E_n}, V_{E_n}) = \int_{T'_E = s + \frac{2n}{\lambda r} + f + e}^T e^{-\frac{\rho}{12}t} dt.$$

The expected value of efficacy and its variance when the stratified and adaptive enrichment designs are successful are calculated by simulation. The expected discounted total revenue for sample size n in each arm of the stratified design (R_{S_n}) is given by

$$\begin{aligned} R_{S_n} &= PoS_{S_{n,1}} \cdot M \cdot m(E_{S_{n,1}}, V_{S_{n,1}}) \int_{T'_S = s + \frac{2n}{r} + f + e}^T e^{-\frac{\rho}{12}t} dt \\ &\quad + PoS_{S_{n,2}} \cdot \lambda M \cdot m(E_{S_{n,2}}, V_{S_{n,2}}) \int_{T'_S = s + \frac{2n}{r} + f + e}^T e^{-\frac{\rho}{12}t} dt, \end{aligned}$$

where $E_{S_{n,1}}$ and $E_{S_{n,2}}$ are the expected values of efficacy in the case of $S_{S,1}$ and $S_{S,2}$, respectively, and $V_{S_{n,1}}$ and $V_{S_{n,2}}$ are their variances. Additionally, the expected discounted total revenue for sample size n in each arm of the adaptive enrichment design (R_{A_n}) is given by

$$\begin{aligned}
R_{A_n} = & PoS_{A_{1n,1}} \cdot M \cdot m(E_{A_{1n,1}}, V_{A_{1n,1}}) \int_{T'_A = s + \frac{2n}{r} + f + e}^T e^{-\frac{\rho}{12}t} dt \\
& + PoS_{A_{2n,1}} \cdot \lambda M \cdot m(E_{A_{2n,1}}, V_{A_{2n,1}}) \int_{T'_A = s + \frac{2n}{r} + f + e}^T e^{-\frac{\rho}{12}t} dt \\
& + PoS_{A_{n,2}} \cdot M \cdot m(E_{A_{n,2}}, V_{A_{n,2}}) \int_{T'_A = s + \frac{2n}{r} + f + e}^T e^{-\frac{\rho}{12}t} dt \\
& + PoS_{A_{n,3}} \cdot \lambda M \cdot m(E_{A_{n,3}}, V_{A_{n,3}}) \cdot \int_{T'_A = s + \left(\frac{2nt}{r} + \frac{2n-2nt}{\lambda r}\right) + f + e}^T e^{-\frac{\rho}{12}t} dt,
\end{aligned}$$

where $E_{A_{1n,1}}, E_{A_{2n,1}}, E_{A_{n,2}}$ and $E_{A_{n,3}}$ are the expected values of efficacy in the case of $S_{A_{1n,1}}, S_{A_{2n,1}}, S_{A_{n,2}}$ and $S_{A_{n,3}}$, respectively, and $V_{A_{1n,1}}, V_{A_{2n,1}}, V_{A_{n,2}}$ and $V_{A_{n,3}}$ are their variances.

Appendix 3. Expected phase III cost

The discounted phase III cost in the enriched and stratified designs is given below:

$$\begin{aligned}
C_{E_n} = & r\lambda(c_{sub} + c_{bio}) \int_s^{s + \frac{2n}{r\lambda}} e^{-\frac{\rho}{12}t} dt + c_{trial} \int_s^{s + \frac{2n}{r\lambda} + f} e^{-\frac{\rho}{12}t} dt + c_{CDx}, \\
C_{S_n} = & r(c_{sub} + c_{bio}) \int_s^{s + \frac{2n}{r}} e^{-\frac{\rho}{12}t} dt + c_{trial} \int_s^{s + \frac{2n}{r} + f} e^{-\frac{\rho}{12}t} dt + c_{CDx}.
\end{aligned}$$

Additionally, let p_k ($k = 1, \dots, 4$) denote the probability of selecting case 1 to case 4 in Table 1 and $C_{A_n,k}$ denote the cost, excluding the interim analysis cost required in each case. The discounted phase III cost of conducting the adaptive enrichment design (C_{A_n}) is given by the following equation:

$$C_{A_n} = \sum_{k=1}^4 p_k C_{A_n,k} + c_{interim} e^{-\frac{\rho}{12} \left(s + \frac{2nt'}{r} + f \right)} + c_{CDx}.$$

Note that

$$\begin{aligned} C_{A_n,1} &= C_{A_n,2} = r(c_{sub} + c_{bio}) \int_s^{s+\frac{2n}{r}} e^{-\frac{\rho}{12}t} dt + c_{trial} \int_s^{s+\frac{2n}{r}+f} e^{-\frac{\rho}{12}t} dt, \\ C_{A_n,3} &= (c_{sub} + c_{bio}) \left\{ r \int_s^{s+\frac{2nt'}{r}} e^{-\frac{\rho}{12}t} dt + r\lambda \int_{s+\frac{2nt'}{r}}^{s+\frac{2nt'}{r}+\frac{2n-2nt'}{r\lambda}} e^{-\frac{\rho}{12}t} dt \right\} \\ &\quad + c_{trial} \int_s^{s+\frac{2nt'}{r}+\frac{2n-2nt'}{r\lambda}+f} e^{-\frac{\rho}{12}t} dt, \\ C_{A_n,4} &= r(c_{sub} + c_{bio}) \int_s^{s+\frac{2nt'}{r}} e^{-\frac{\rho}{12}t} dt + c_{trial} \int_s^{s+\frac{2nt'}{r}+f} e^{-\frac{\rho}{12}t} dt. \end{aligned}$$

Appendix 4. The value of parameters used in the numerical example and the motivating example

Table A1. The value of variable parameters used in the numerical example

[Insert Table A1]

Table A2. The values of fixed parameters used in the numerical example

[Insert Table A2]

Table A3. The values of parameters used in the motivating example

[Insert Table A3]

Table 1. Decision rule at the interim analysis

	$\hat{\Delta}_{P,1} \geq c_P$	$\hat{\Delta}_{P,1} < c_P$
$\hat{\Delta}_{F,1} \geq c_F$	<u>Case 1</u> Continue with F and P	<u>Case 2</u> Continue with F only
$\hat{\Delta}_{F,1} < c_F$	<u>Case 3</u> Continue with P only	<u>Case 4</u> Stop for futility

Table 2. Scenarios of efficacy in historical data

Historical data	Scenario			
	1	2	3	4
Mean treatment effect in P ($\hat{\Delta}_P^1$)	0.4	0.4	0.4	0.4
Mean treatment effect in N ($\hat{\Delta}_N^1$)	0.4	0.2	0.0	-0.2
Sample size in P per arm (n_P^1)	100 λ			
Sample size in N (n_N^1)	100(1- λ)			

Table 3. Result of clinical trial in the motivating example

Historical data	ApoE4 status	Sample size ^{a)}		Result	Estimates of variance of treatment effect
		Arm E	Arm P		
Risner et al. ¹¹	Positive	36	35	1.49 [-0.81 , 3.80]	1.34
	Negative	42	43	-2.94 [-5.06 , -0.81]	1.15

a) Arm E and Arm P denote an experimental arm and a placebo arm, respectively. The experimental arms are rosiglitazone and donepezil in Risner et al.¹¹, respectively.

Table 4. Predictive distribution of treatment effects in each population

Populations	$\tau=0.25$	$\tau=0.99$
ApoE4-positive	N(1.49 , 1.59)	N(1.49 , 5.28)
ApoE4-negative	N(-2.94 , 1.40)	N(-2.94 , 5.09)
Full	N(-0.06, 0.84)	N(-0.06, 2.85)

Table 5. The optimal sample size, PoS, and eNPV in the motivating example

τ	Designs	Optimal sample size (total)	PoS	eNPV
0.25	Traditional	N/A	N/A	N/A
	Enriched	80	0.74	1538
	Stratified	310	0.77	1658
	Adaptive Enrichment	180	0.80	1721
0.99	Traditional	N/A	N/A	N/A
	Enriched	130	0.71	606
	Stratified	370	0.70	1161
	Adaptive Enrichment	190	0.70	1430

N/A indicates that the PoS of the design did not exceed 0.70 in the maximum sample size.

Table A1. The value of variable parameters used in the numerical example

Parameters	Setting value
Prevalence of P among F (λ)	0.2, 0.5, 0.8
Between-trial standard deviation of efficacy (τ)	0.05, 0.2
Threshold value of PoS required to go to phase III (c_{PoS})	0.7, 0.8

Table A2. The values of fixed parameters used in the numerical example

Parameters	Setting value
Design and cost parameters of phase III (common)	
Level of significance (α , two-sided)	0.05
Standard deviation of response (σ)	1
Enrollment speed in full population (r , subjects/month)	40
Period from planning phase III to first patient in (s , month)	3
Follow up period (f , month)	6
Minimal clinically significant difference in efficacy (Δ_0)	0
Patient cost ($c_{patient}$, MUSD)	0.08
Trial cost (c_{trial} , MUSD)	3
Cost of evaluating biomarker (c_{bio} , MUSD)	0.001
Development cost of companion diagnostic (c_{cdx} , MUSD)	10
Examination period (e , month)	18
Design and cost parameters of phase III (only for adaptive enrichment design)	

Criteria of population selection in the interim analysis (c_F , c_P)	0.1 (common)
Timing of interim analysis (t)	0.5
Cost of conducting the interim analysis ($c_{interim}$, MUSD)	1.5
Financial parameter	
The rest of the patent life at the time of planning phase III design (T , month)	120
Discount rate (ρ , /year)	0.1
Efficacy that some patients want to switch to new drugs (A)	0.1
Efficacy that all patients want to switch to new drugs (B)	0.5

Table A3. The values of parameters used in the motivating example

Parameters	Setting value
Threshold value of PoS required to go to phase III (c_{PoS})	0.7
Cost of evaluating biomarker (c_{bio} , MUSD)	0.00015
Development cost of companion diagnostic (c_{cdx} , MUSD) ^{a)}	0
Market size (M , MUSD/year)	3000
Criteria of population selection in the interim analysis (c_F , c_P)	0.5 (common)
The mean treatment effect that some patients consider switching to new drugs (A)	0.5
The mean treatment effect that all patients consider switching to new drugs (B) ^{b)}	4

Note that the values of parameters included in Table 1A but not in Table 2A are the same as those in Table A1.

a) No need to develop companion diagnostic because ApoE genetic test kits are on the market.

b) Raina et al.²⁹ suggested that a change of 4 points of ADAS-cog from baseline is considered clinically significant for patients with mild to moderate dementia.

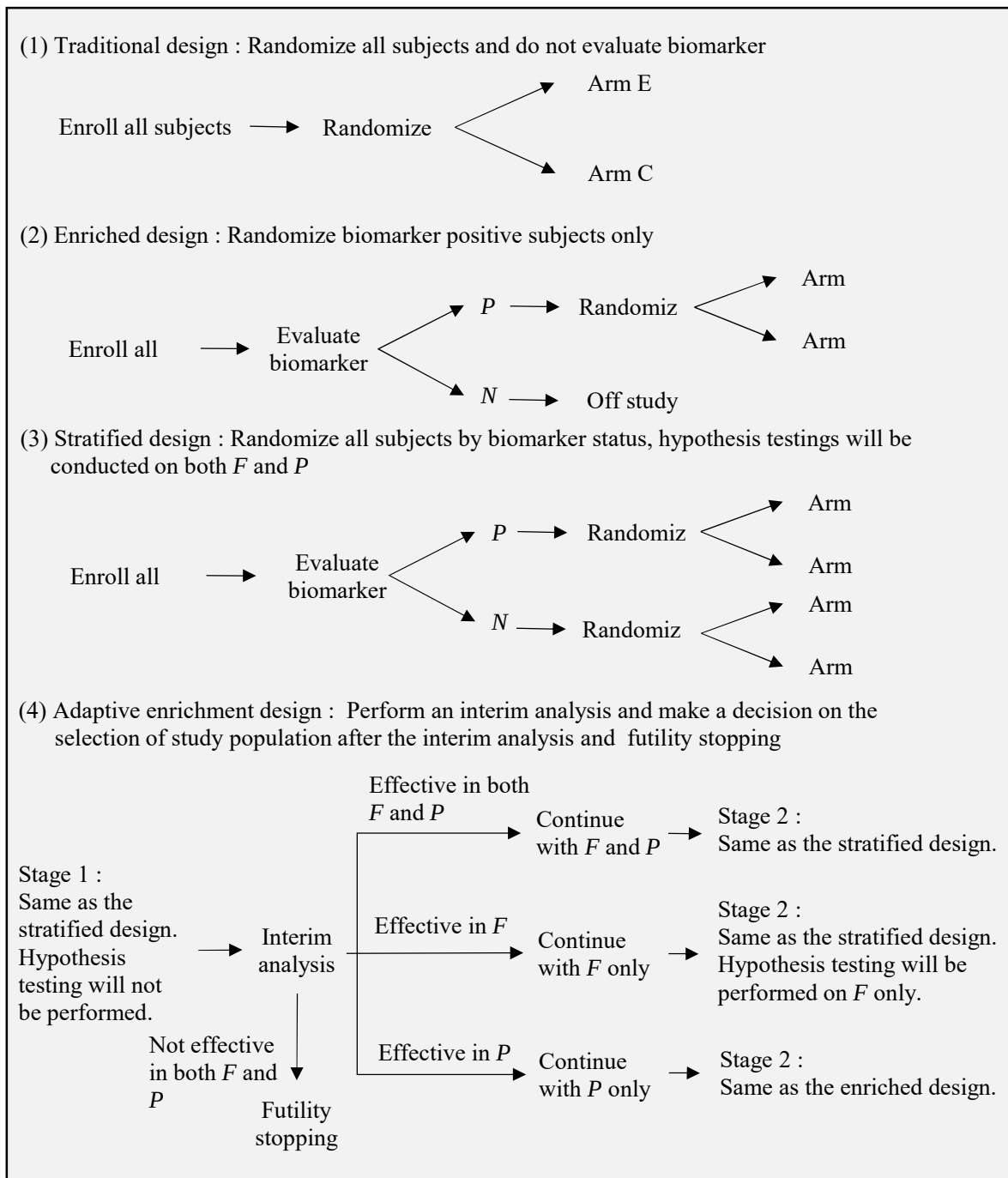


Figure 1. Outline of considered study designs.

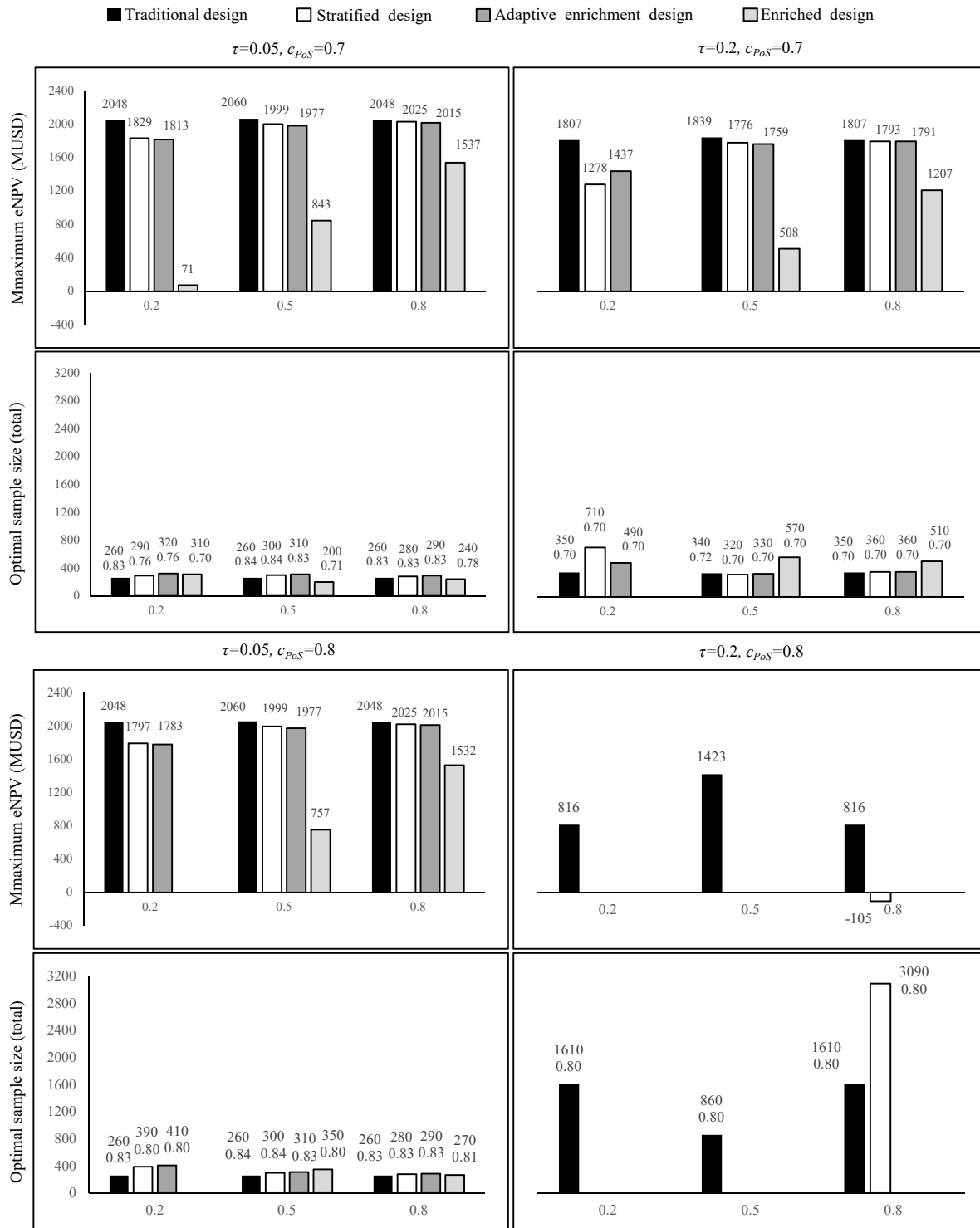


Figure 2a. Maximum eNPV (MUSD) and optimal sample size (total) of each design in scenario 1. The horizontal axis represents λ . The upper number in the optimal sample size figure is the optimal sample size (total), and the lower number is PoS. Designs that did not result in $PoS \geq c_{PoS}$ at the maximum sample size were not outputted.

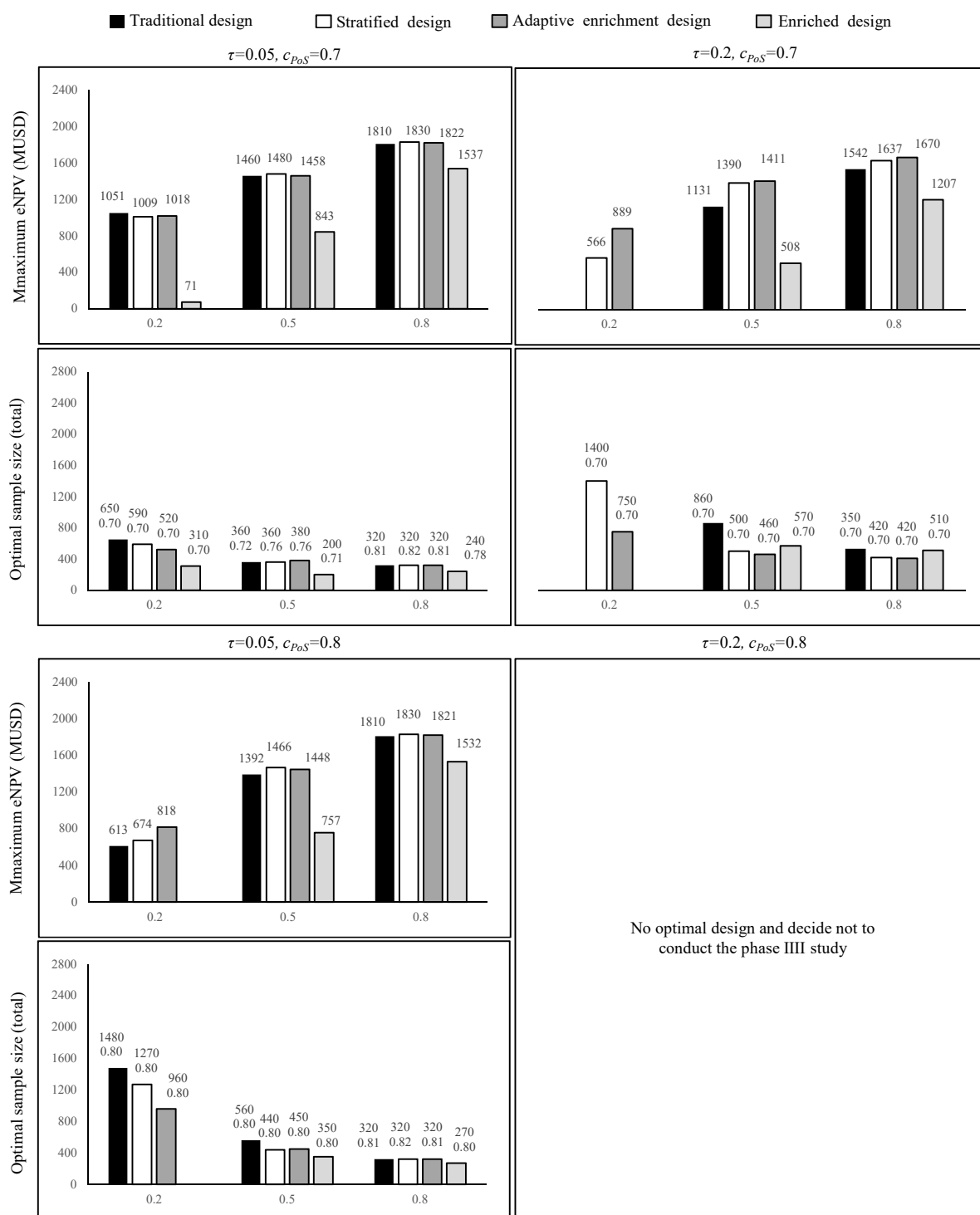


Figure 2b. Maximum eNPV (MUSD) and optimal sample size (total) of each design in scenario 2. See the explanatory note in Figure 2a.

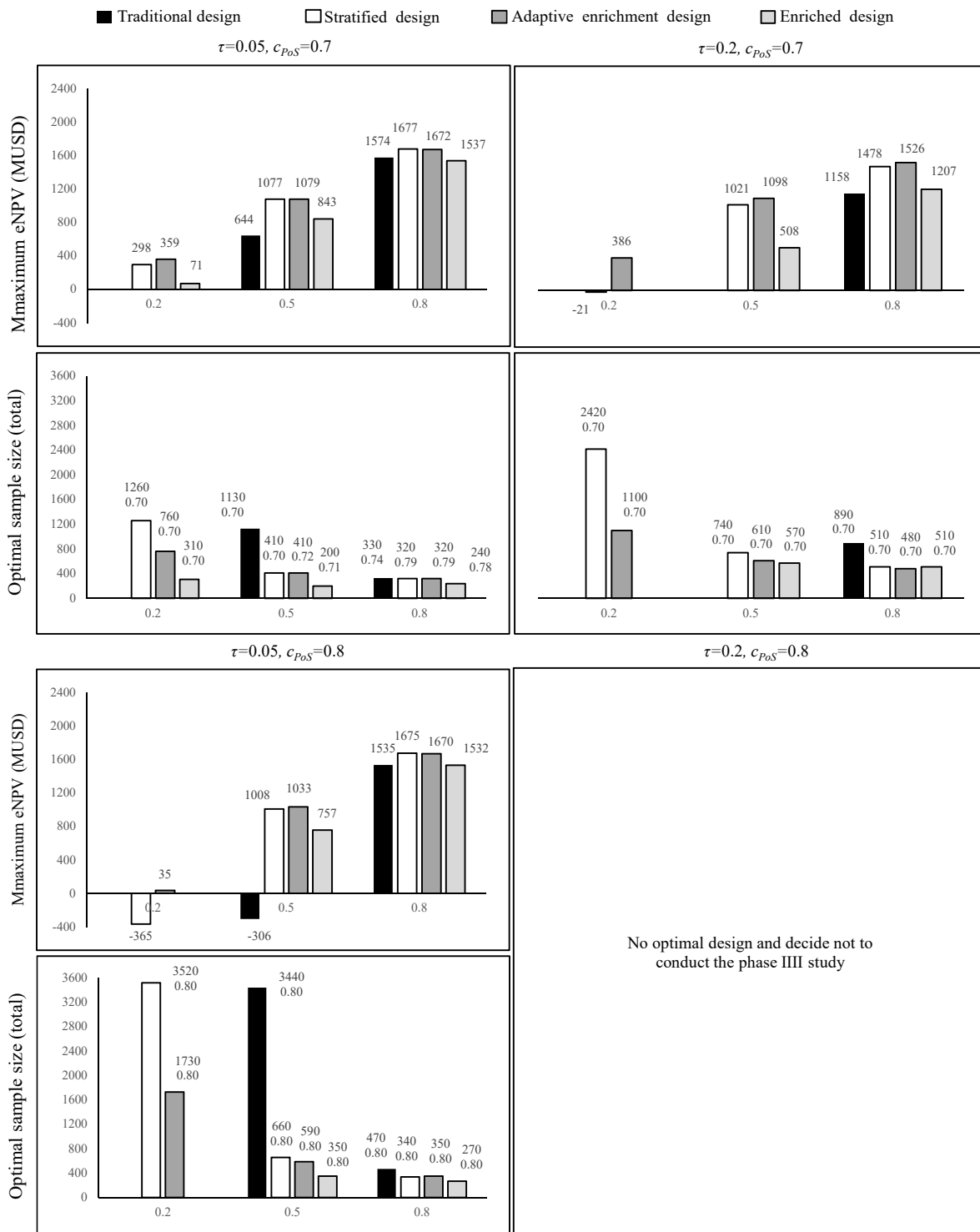


Figure 2c. Maximum eNPV (MUSD) and optimal sample size (total) of each design in scenario 2. See the explanatory note in Figure 2a.

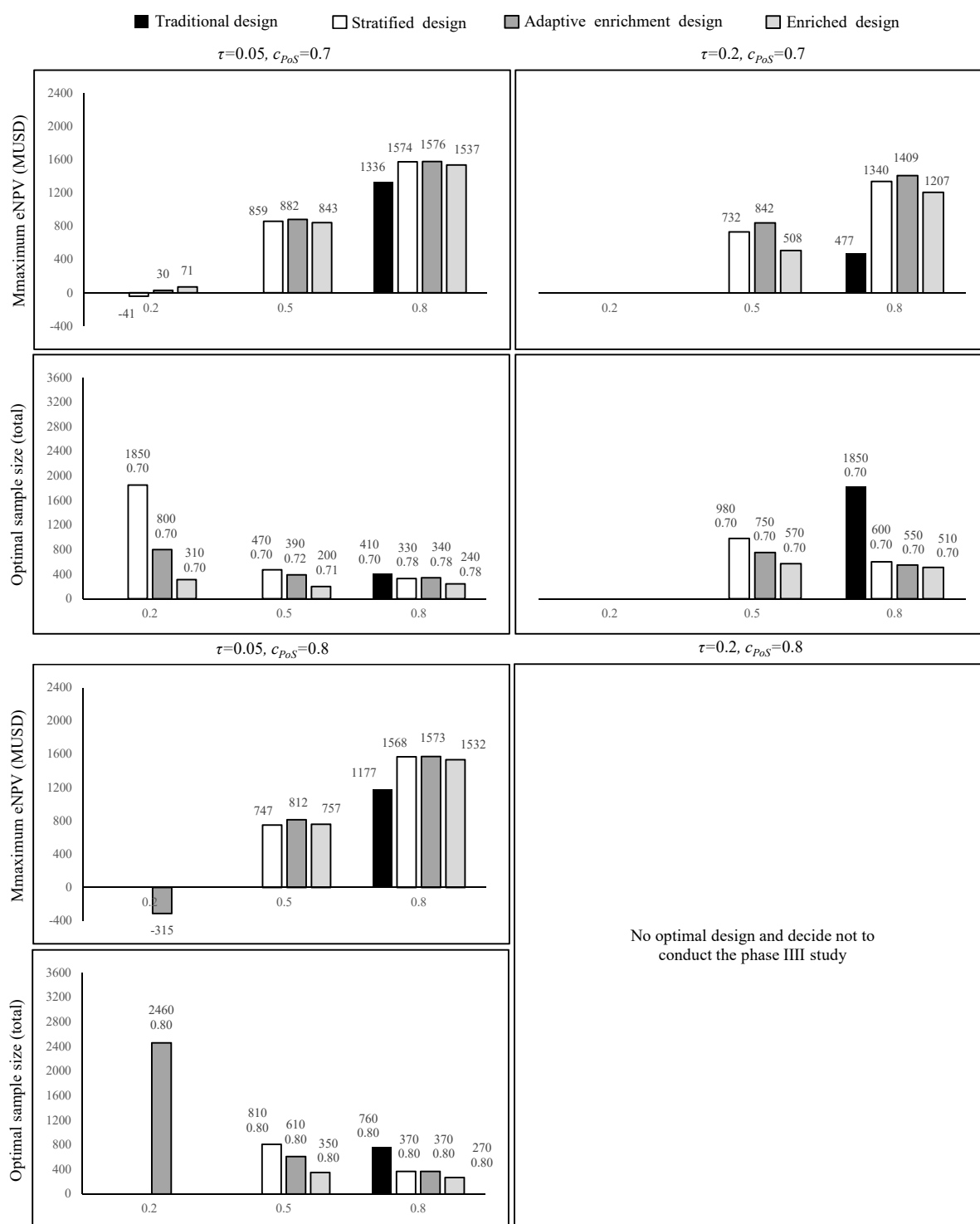


Figure 2d. Maximum eNPV (MUSD) and optimal sample size (total) of each design in scenario 2. See the explanatory note in Figure 2a.

Supplementary material to Optimal decision criteria for the study design and sample size of a biomarker-driven phase III trial

Akira Takazawa MSc.^{1,2}, Satoshi Morita PhD.²

Data Science Department, ONO Pharmaceutical Co., Ltd. Osaka, Japan¹,

Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate

School of Medicine²

Details of the considered stratified design

Z-statistic, which is used to test two hypotheses $H_{0,F}: \Delta_F \leq 0, H_{1,F}: \Delta_F > 0$ and $H_{0,P}: \Delta_P \leq 0, H_{1,P}: \Delta_P > 0$, is given by z_F and z_P in the traditional and enriched designs, respectively.

The corresponding p-values is given by $p_F = 1 - \Phi(z_F)$ and $p_P = 1 - \Phi(z_P)$,

respectively. In Spiessens and Debois test, the interaction hypothesis $H_{0,\{F,P\}} = H_{0,F} \cap H_{0,P}$

is rejected when $p_F \leq \alpha_1$ or $p_P \leq \alpha_2$, and $H_{0,F}$ and $H_{0,P}$ are rejected when $p_F \leq \alpha$ and

$p_P \leq \alpha$, respectively if the interaction hypothesis is rejected. Note that α_1 and α_2 should

be set to satisfy the following equation so that the Family Wise Type Error Rate is less than

or equal to α . Here, we set $\alpha_1 = \alpha_2$.

$$P(p_F \leq \alpha_1 \text{ or } p_P \leq \alpha_2 | H_{0,\{F,P\}}) = \alpha.$$

We judge that the study is successful in F when $\{p_F \leq \alpha' \cup p_P \leq \alpha'\} \cap \{p_F \leq \alpha\} \cap$

$\{\hat{\Delta}_P \geq \Delta_0 \cap \hat{\Delta}_N \geq \Delta_0\}$ and in P when $\{p_F \leq \alpha' \cup p_P \leq \alpha'\} \cap \{p_P \leq \alpha\}$.

Details of the considered adaptive enrichment design

The period before the interim analysis is defined as stage 1 and that after the interim analysis as stage 2. n_j is the sample size in stage j ($j=1,2$), and the sample size in each arm after the interim analysis is given by $n_2 = n - nt'$. In case 1 in the Table 1, in order to test F and P , the test is performed using the Spiessens and Debois method, as in the stratified design. In case 2 and case 3, only F or P will be continued, thus n_2 subjects in each arm will be enrolled in stage 2.

Similar to the stratified design, hypotheses in the adaptive enrichment design are set as $H_{0,F}$ and $H_{0,P}$, and the interaction hypothesis is $H_{0,\{F,P\}} = H_{0,F} \cap H_{0,P}$. We set the p-value calculated from stage j data for testing each hypothesis as $p_{F,j}$, $p_{P,j}$, and $p_{\{F,P\},j}$ respectively. Let the significance level of F and P used for the interaction hypothesis test be the same value and the test statistics in population i of stage j be $z_{i,j}$ and $z_{max,j} = \max\{z_{F,j}, z_{P,j}\}$. The p-value in stage j to test interaction hypothesis is given by:

$$1 - \int_{-\infty}^{z_{max,j}} \Phi\left(\frac{z_{max,j} - \sqrt{\lambda}z}{\sqrt{1-\lambda}}\right) \varphi(z) dz.$$

Bauer and Kohne¹ proposed using the inverse normal combination function as a method to integrate p-values obtained from different stages. Because $w_j = \sqrt{n_j/n}$ and $n_1 = nt'$, $n_2 = n - nt'$, the combination function becomes $C(p_1, p_2) = 1 - \Phi(\sqrt{t'}z_{p_1} + \sqrt{1-t'}z_{p_2})$, and the null hypothesis is rejected when $C(p_1, p_2) \leq \alpha$.

Based on the information provided above Table S1 shows the criteria to decide that the study was successful in each case in Table 1. Note that $S_{A1,1}$, $S_{A2,1}$, $S_{A,2}$ and $S_{A,3}$ indicate the situation in which the study is successful in each case, and $S_{A,4}$ indicates the situation in which the study stopped for futility at the interim analysis.

Table S1 : Criteria to decide that the study was successful in each case

Case 1 :

$S_{A1,1}$: Success in F if

$$\{C(p_{\{F,P\},1}, p_{\{F,P\},2}) \leq \alpha\} \cap \{C(p_{F,1}, p_{F,2}) \leq \alpha\} \cap \{\hat{\Delta}_P \geq \Delta_0 \cap \hat{\Delta}_N \geq \Delta_0\}$$

$S_{A2,1}$: Success in P if $\{C(p_{\{F,P\},1}, p_{\{F,P\},2}) \leq \alpha\} \cap \{C(p_{P,1}, p_{P,2}) \leq \alpha\}$

Case 2 :

$S_{A,2}$: Success in F if

$$\{C(p_{\{F,P\},1}, p_{F,2}) \leq \alpha\} \cap \{C(p_{F,1}, p_{F,2}) \leq \alpha\} \cap \{\hat{\Delta}_P \geq \Delta_0 \cap \hat{\Delta}_N \geq \Delta_0\}$$

Case 3 :

$S_{A,3}$: Success in P if $\{C(p_{\{F,P\},1}, p_{P,2}) \leq \alpha\} \cap \{C(p_{P,1}, p_{P,2}) \leq \alpha\}$

Reference

1. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses.

Biometrics. 1994;50:1029-1041.

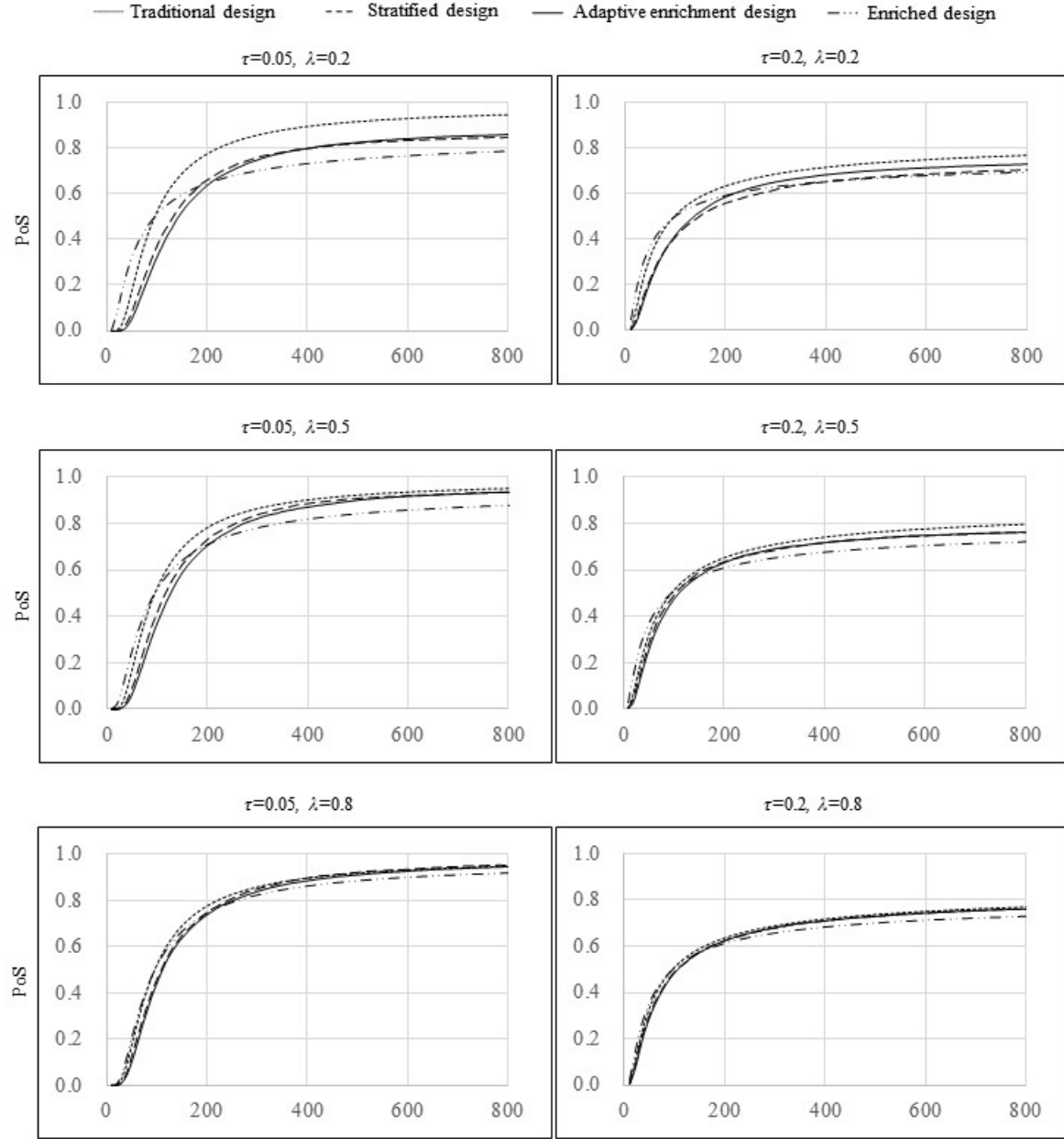


Figure S1. Sample size (total) and PoS of each design in scenario 1. The horizontal axis represents the sample size (total).

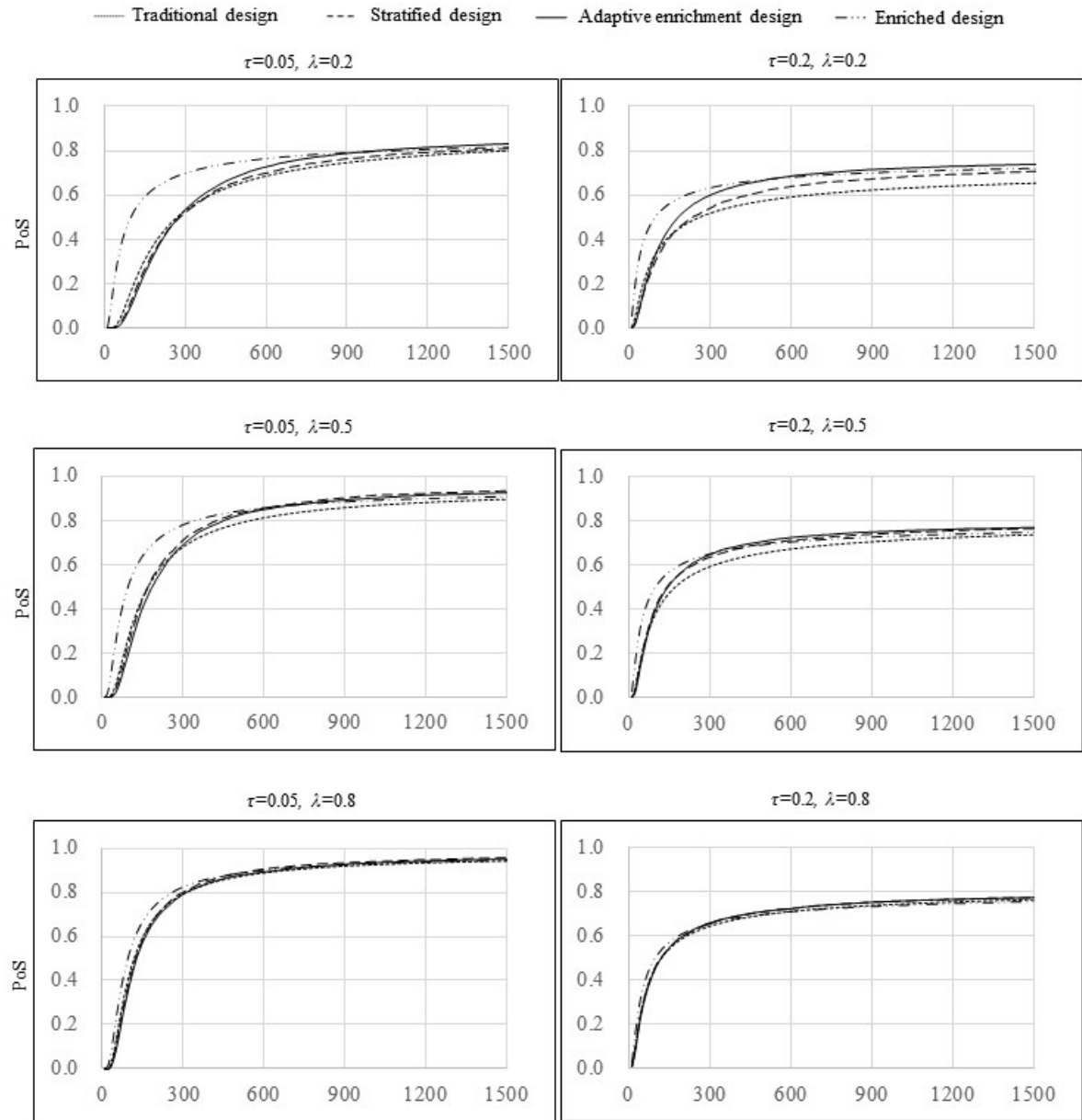


Figure S2. Sample size (total) and PoS of each design in scenario 2. The horizontal axis represents the sample size (total).

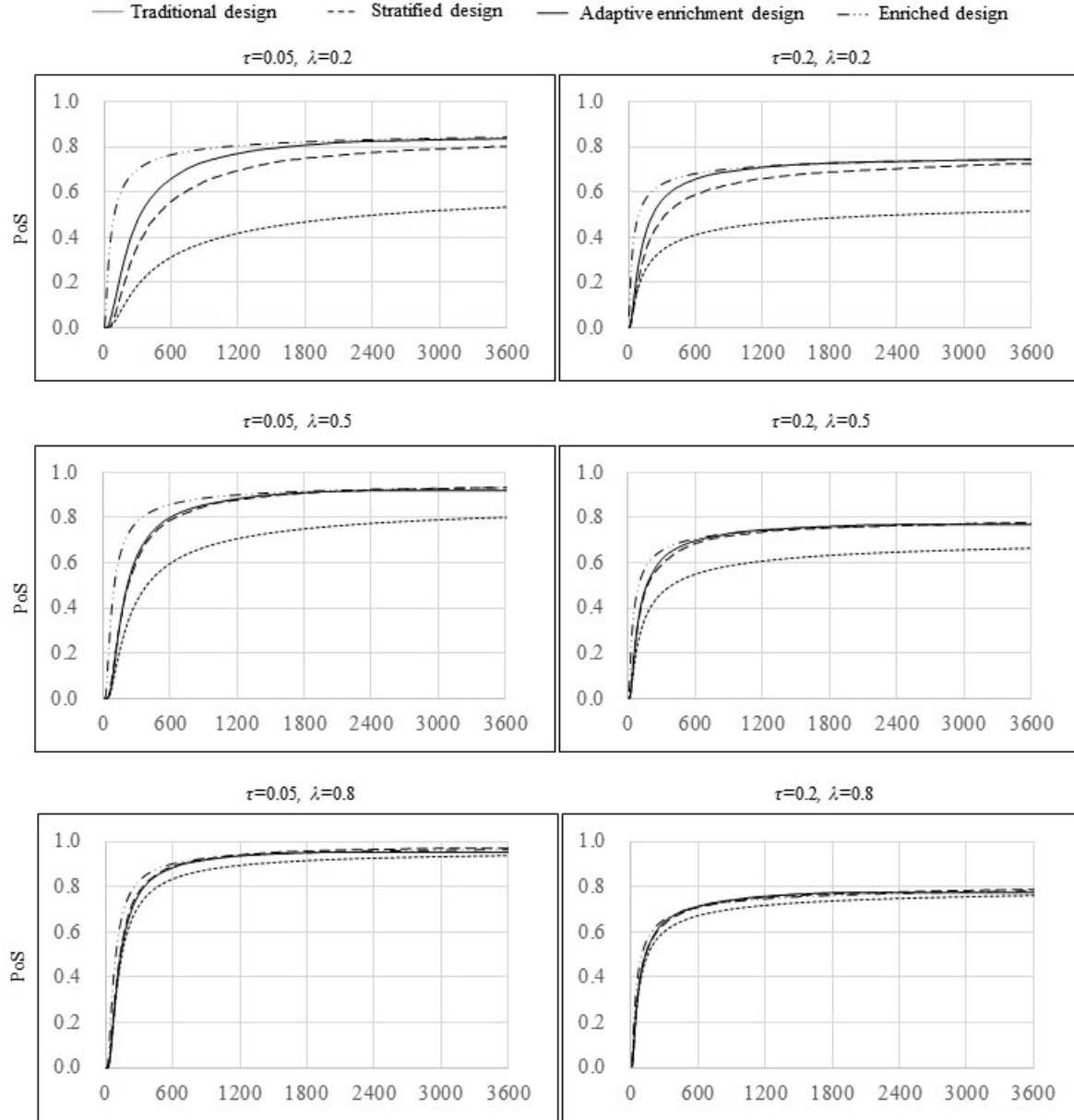


Figure S3. Sample size (total) and PoS of each design in scenario 3. The horizontal axis represents the sample size (total).

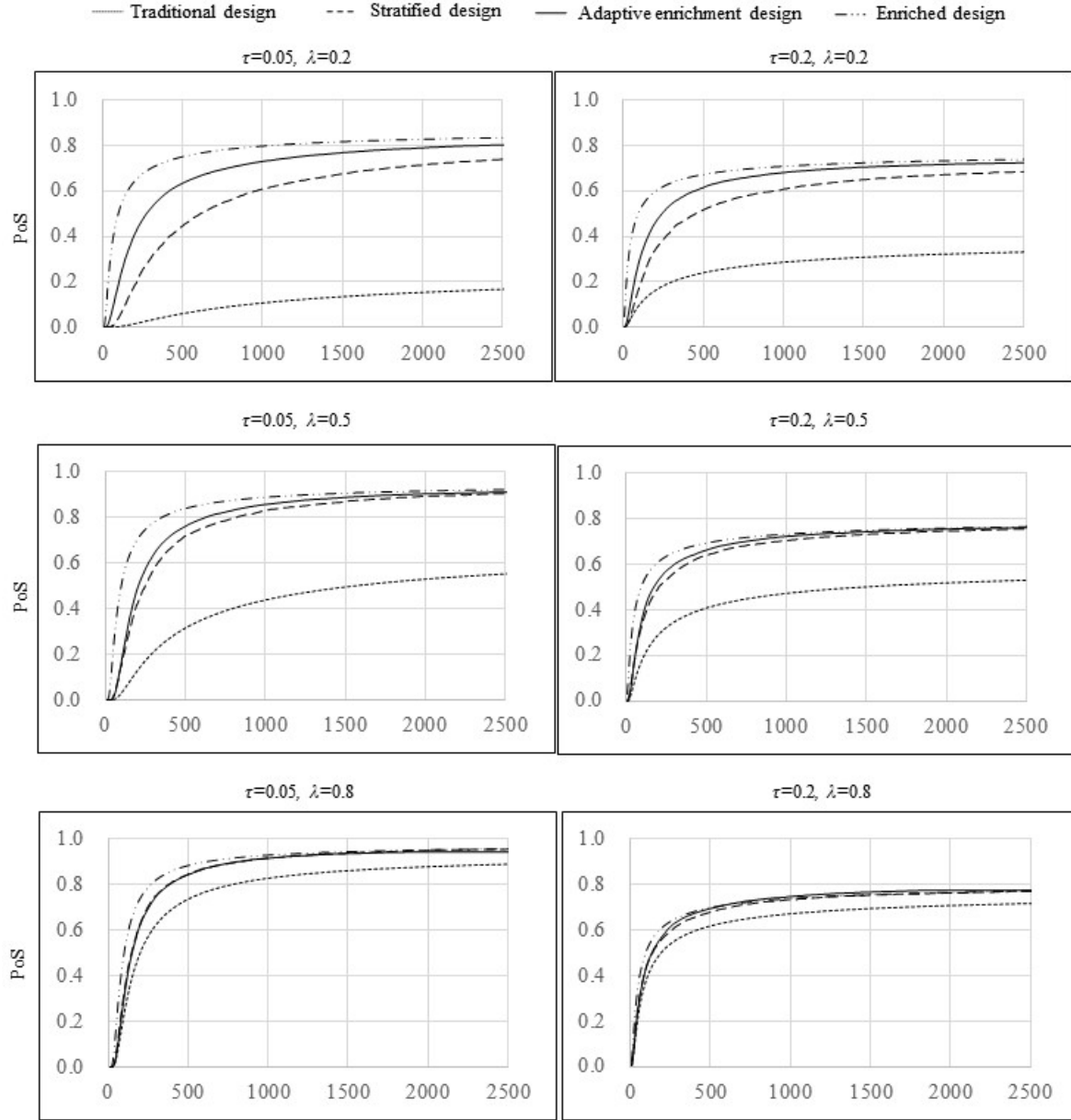


Figure S4. Sample size (total) and PoS of each design in scenario 4. The horizontal axis represents the sample size (total).