Journal of Statistical Science and Application, February 2017, Vol. 5, No. 1-2, 1-15
doi: 10.17265/2328-224X/2017.0102.001

Utility of Adaptive Sample Size Designs and A Review Example

Kohei Uemura[1]
Shin-kasumigaseki Bldg. 3-3-2, Kasumigaseki, Chiyoda-ku, Tokyo 100-0013, Japan
Phone : +81-3-3506-9487, Facsimile: +81-3-3506-9567, E-mail: uemura-kohei@pmda.go.jp
Affiliation: Biostatistics Group, Center for Product Evaluation, Pharmaceuticals and Medical Devices Agency, Tokyo, Japan

Yuki Ando[2]
Address: Shin-kasumigaseki Bldg. 3-3-2, Kasumigaseki, Chiyoda-ku, Tokyo 100-0013, Japan
Phone: +81-3-3506-9448, Facsimile: +81-3-3506-9450, E-mail: ando-yuki@pmda.go.jp
Affiliation: Biostatistics Group, Center for Product Evaluation, Pharmaceuticals and Medical Devices Agency, Tokyo, Japan

Yutaka Matsuyama[3]
Address: 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Phone: +81-3-5841-3520, Facsimile: +81-3-3814-2779, E-mail: matuyama@epistat.m.u-tokyo.ac.jp
Affiliation: Department of Biostatistics, School of Public Health, University of Tokyo, Tokyo, Japan

Sample size can be a key design feature that not only affects the probability of a trial's success but also determines the duration and feasibility of a trial. If an investigational drug is expected to be effective and address unmet medical needs of an orphan disease, where the accrual period may require many years with a large sample size to detect a minimal clinically relevant treatment effect, a minimum sample size may be set to maintain nominal power. In limited situations such as this, there may be a need for flexibility in the initial and final sample sizes; thus, it is useful to consider the utility of adaptive sample size designs that use sample size re-estimation or group sequential design. In this paper, we propose a new adaptive performance measure to consider the utility of an adaptive sample size design in a trial simulation. Considering that previously proposed sample size re-estimation methods do not take into account errors in estimation based on interim results, we propose Bayesian sample size re-estimation criteria that take into account prior information on treatment effect, and then, we assess its operating characteristics in a simulation study. We also present a review example of sample size re-estimation mainly based on published paper and review report in Pharmaceuticals and Medical Devices Agency (PMDA).

Keywords: adaptive design, sample size re-estimation, group sequential design, utility, Bayesian method, regulatory discussion, review example

Introduction

Recently, it has become more challenging to cope with a low probability of success in clinical trials because of the high costs of new drug development. An adaptive design is considered a promising tool for efficient drug development. This paper focuses on two adaptive sample size designs, classical group sequential design (GSD) and sample size re-estimation (SSR), with adaptations that are simple and basic. The sample size

Disclaimer of Author 1 and 2: The views expressed in this article are those of the authors alone and do not necessarily represent the official views of the Pharmaceuticals and Medical Devices Agency.

is a key design feature that not only affects a trial's probability of success but also determines the duration and feasibility of a trial and the entire period of drug development. On the other hand, adaptive sample size designs require additional resources to construct an appropriate system to conduct interim decision making for adaptations and to maintain the integrity of a trial. Therefore, it is important to assess 'Why use adaptive sample sizes?' and to assess whether adaptive sample size designs can be effective in a practical manner.

Statistical discussion points related to adaptive sample size designs include inflation of the type I error rate and statistical bias. GSD can control the type I error rate; however, a trial could be terminated early because it appears beneficial owing to an overestimate of the treatment effect. It should be noted that an overestimation of an interim result cannot be corrected by type I error control. Thus, uncertainty about the magnitude of a treatment effect can be concerned with GSD, as discussed in the review of Sunitinib for pancreatic neuroendocrine tumors (U.S. Food and Drug Administration, FDA, 2011). SSR can cope with alpha inflation and statistical bias using previously proposed methods (Bauer and Kohne, 1999; Proschan and Hunsberger, 1995; Cui et al, 1999; Chen et al, 2004; Uemura et al, 2008; Brannath et al, 2006; Liu, Proschan and Pledger (2002)). Note that SSR with an early efficacy stopping rule can have the same overestimate problem as GSD. Another statistical discussion point of adaptive sample size designs is that the statistical efficiency of such designs is still controversial. While GSD can be more efficient than SSR (Jennison and Turnbull, 2003; Tsiatis and Mehta, 2003), simulation settings by Jennison and Turnbull are not easy to apply in practice (Hung et al., 2005). A related problem, the statistical inefficiency of SSR, occurs because the interim estimate of a treatment effect can vary and may be unreliable. This indicates that there is some risk that re-estimated sample sizes might be erroneous.

Practical and regulatory discussion points of adaptive sample size designs include the potential risk of operational bias with SSR, which cannot be fully avoided and statistically corrected if it occurs. Decision of sample size escalation and/or re-estimatited sample size can be more informative than an interim decision of non-stop in GSD. The FDA draft guidance for adaptive design classifies SSR as a less well-understood design (FDA, 2010). It is necessary to carefully consider its application in a confirmatory trial. Note, however, that operational bias can occur with GSD. If an unplanned additional interim analysis is conducted and real-time monitoring indicates a promising trend, an unplanned interim stop could seriously inflate the type I error rate and introduce a bias toward an overestimation of the treatment effect. Another practical and regulatory discussion point is that the practical motivation and utility of SSR needs to be explained and justified, considering the potential risk of operational bias with this design. In general, the need for flexibility in initial and final sample sizes is not clear because the feasibility of a trial and the risk of underpowering a study may leave little room for flexibility in sample sizes. For example, a maximally feasible and/or empirically reasonable sample size may be set for a fixed design without interim analysis and re-estimation of sample size. With limited resources, the gap between these sample sizes may be smaller, especially in Japan. On the other hand, if an investigational drug is expected to be effective and to meet unmet medical needs for an orphan disease (e.g., biologics targeting an inflammatory cytokine that is strongly linked to disease activity, based on evidence from a similar disease) but the accrual period may require many years with a large sample size to detect a minimal clinically relevant treatment effect, a minimum sample size may be set to maintain nominal power for an expected treatment effect to meet the unmet need as soon as possible. In limited situations such as this, there may be the need for flexibility in initial and final sample sizes, and it may be useful to consider the practical valueof SSR.

In this paper, we propose a new adaptive performance measure that takes into consideration the utility of an adaptive sample size design in a trial simulation. In addition, as mentioned above, previously proposed sample size re-estimation methods do not take into account errors in estimation based on an interim result. Therefore, we propose Bayesian sample size re-estimation criteria that use prior information on treatment effects, and we assess its operating characteristics in a simulation study.In addition, we present a review example of sample size re-estimation mainly based on published paper and review report in PMDA.

A New Adaptive Performance Measure

We will consider a placebo-controlled randomized two-arm clinical trial. Let Y_{ij} denote the primary endpoint for subject i allocated to group j ($j = 1\text{: investigational drug}, 2\text{: placebo}$) and follow a normal distribution with a mean of $\mu_1 = \delta, \mu_2 = 0$ and a common variance of $\sigma^2 = 1,$ without the loss of generality. The null hypothesis is $H_0\colon \delta = 0$ and the alternative hypothesis is $H_1\colon \delta > 0$. The pre-assumed effect size is δ_{pre}, nominal type I error rate is α, nominal power is $1-\beta,$ and the initial planned sample size $N_{initial}$ per group is set as follows:

$$N_{initial} = 2\left(\frac{z_\alpha + z_\beta}{\delta_{pre}}\right)^2, \tag{1}$$

where z_u is the $(1-u)th$ quantile of a standard normal distribution. For simplicity, one interim analysis is planned at information time t $(0 < t < 1)$. Let stage 1 denote data fixed before the interim analysis and stage 2 denote data fixed after t until the final analysis. Test statistic $Z_1,$ based on stage 1, is calculated as follows:

$$Z_1 = \frac{1}{\sqrt{2n_1}}\sum_{i=1}^{n_1}(Y_{i1} - Y_{i2}), \tag{2}$$

where n_1 is the stage 1 sample size. The final determined trial sample size N_{final} is as follows:

$$N_{final} = \begin{cases} n_1, & if\; Z_1 \leq 0 \text{ or } Z_1 \geq c_1 \\ N_{re-estimated}, & if 0 < Z_1 < c_1 \end{cases}, \tag{3}$$

where the efficacy and futility stopping boundaries for the interim analysis are respectively $c_1, 0$ and $N_{re-estimated}$ is the trial sample size re-estimated based on an interim estimate of effect size $\hat{\delta}_1$. The final analysis is based on $Z_{weig\,hted}$, which is the weighted Z statistic proposed by Cui, Hung, and Wang (1999) in a group sequential trial setting.

$$Z_{weig\,hted} = \sqrt{t}Z_1 + \sqrt{1-t}Z_2^*, \tag{4}$$

where Z_2^* denotes the same Z statistic as equation (2), with subjects of $n_1 + 1 \sim N_{re-estimated}$.

In this section, we propose a new adaptive performance measure that takes into account the utility of an adaptive sample size design in a trial simulation. Note that we will mainly assume the limited situation of a promising orphan drug development because the need for flexibility in initial and final sample sizes may not be clear in other situations, as described in the introduction. Under a promising orphan drug situation with a long period of patient recruitment (e.g., several years), we think it may be very important to minimize the sample size and not to increase the sample size unnecessarily to eliminate a long waiting time of severely diseased patients who may benefit from a new drug approval. Then, we will consider the opposite SSR situation, where one can expect a much larger effect size than the minimal clinically relevant effect compared to the usual SSR situation with small pre-assumed effect size that suppose to extend the maximum sample size if there remains some chance to show a minimal clinically relevant treatment effect.

For the first step, we defined the expected over-size (EOS) as follows:

$$EOS(\delta) = \{[ASN(\delta) - N_{ideal}(\delta)]_{+} / N_{ideal}(\delta)\} \times 100, \tag{5}$$

where $ASN(\delta)$ is the average sample number that is expectation of the final sample size under an adaptive sample size design and true effect size δ in the long run, $[.]_{+}$ denotes a function that includes a value within a square bracket only if $[.] > 0$, otherwise, the value is considered 0. $N_{ideal}(\delta)$ is the ideal sample size and is calculated based on the true effect size δ as follows:

$$N_{ideal}(\delta) = 2\left(\frac{z_{\alpha} + z_{\beta}}{\delta}\right)^{2}. \tag{6}$$

For example, if $N_{ideal}(\delta)$ is 300 and requires an accrual time of 3 years, $EOS(\delta) = 50\%$ means an adaptive sample size trial tends to reach 450 patients, and the accrual timeis 4.5 years on average in the long run. For the second step, we defined the expected under-power (EUP) as follows:

$$EUP(\delta) = \left\{\left[2\left(\frac{z_{\alpha} + z_{1-POWER(\delta)}}{\delta}\right)^{2} - N_{ideal}(\delta)\right]_{-} \Big/ \left[2\left(\frac{z_{\alpha} + z_{0.5}}{\delta}\right)^{2} - N_{ideal}(\delta)\right]\right\} \times 100, \tag{7}$$

Where $POWER(\delta)$ is the probability that is expectation of a final test result that indicates 1 with statistical significance; otherwise, it is 0 under an adaptive sample size design with the true effect size δ in the long run. $EUP(\delta)$ is a measure that indicates the extent of under-power in scale of sample size relative to a reference. We selected the reference under-power case of $POWER(\delta) = 0.5$, which means an adaptive sample size design will be expected to provide a marginal result such as a p-value equal to 0.05. One should not consider the under-power case of $POWER(\delta) < 0.5$ as the reference, because an adaptive sample size design should not be a way to rescue a study from poor initial planning and a promising orphan drug is expected to have at least a moderate treatment effect. For the third step, we defined the expected adaptive performance measure (EAP) as follows:

$$EAP(\delta) = EOS(\delta) + EUP(\delta). \tag{8}$$

If one shifted the balance between under-power and over-size toward the riskof under-powering a study, one should set higher values than 0.5 for the reference under-power. $EAP(\delta)$ is equivalent to the adaptive performance score that is shown as a regret function in Liu, Shu, and Cui (2008).

Then, we proposed a new conditional adaptive performance measure (CAP) for which over-size and under-power are evaluated conditionally, given as an interim result. Considering that an adaptive sample size design not only aims to adjust the initial sample size toward the accurate size under true δ in the long run but also aims to minimize the mean error of an adjusted sample size in terms of conditional power given a variable result in the interim analysis, we used a positive or negative function denoted by $[.]_{+}$ or $[.]_{-}$, to account for variability in the difference from an exact adjustment. Note that $[.]_{-}$ includes a value only if $[.] < 0$. Along with $EAP(\delta)$ in the first step, we defined conditional over-size (COS) as follows:

$$COS(\delta) = \frac{\sum_{s=1}^{S}\left[n_{2final\_s} - n_{2ideal\_s}\right]_{+}}{S}, \tag{9}$$

where, S and the subscript s, respectively, denote the number of replications of a trial simulation and the *sth* simulated clinical trial, and n_{2final_s} and n_{2ideal_s}, respectively, denote the finally determined stage 2 sample size, which equals $\left(N_{final\_s} - n_1\right)$, and the ideal stage 2 sample size based on the true δ. Note that both of n_{2final_s} and $\left[n_{2final\_s} - n_{2ideal\_s}\right]_+$ becomes 0 if the *sth* simulated interim data meets the efficacy or futility stopping rule, as in equation (3) and $COS(\delta)$, is on the same scale as the sample size, which is not yet divided by the n_{2ideal_s}. We defined the conditional power function as follows:

$$\begin{aligned} CP\left(Z_{1s}, t, n_{2fina\,l_s};\ \delta_2 = \delta\right) &= Pr\left(Z_{weig\,htd} > c_2 \middle| Z_{1s}, t, n_{2fina\,l_s};\ \delta_2 = \delta\right) \\ &= 1 - \Phi\left(\frac{c_2 - \sqrt{t}Z_{1s}}{\sqrt{1-t}} - \delta\sqrt{\frac{n_{2final\_s}}{2}}\right), \end{aligned} \tag{10}$$

where $\Phi(.)$ denotes a standard normal distribution function.

For the second step, we defined the conditional under-power (CUP) as follows:

$$CUP(\delta) = -\frac{\sum_{s=1}^{S}\left[CP\left(Z_{1s}, t, n_{2fina\,l_s};\ \delta_2 = \delta\right) - (1-\beta)\right]_-}{S}. \tag{11}$$

Note that $\left[CP\left(Z_{1s}, t, n_{2final\_s};\ \delta_2 = \delta\right) - (1-\beta)\right]_-$ takes 0 if the *sth* simulated interim data meets the efficacy stopping rule and takes $1-\beta$ if it meets futility and $CUP(\delta)$ is on the same scale a probability that is not yet on the sample size scale divided by the reference case. Then, n_{2ideal_s} is calculated by solving the equation of $CP\left(Z_{1s}, t, n_{2ideal\_s};\ \delta_2 = \delta\right) = 1 - \beta$ as follows:

$$n_{2ideal\_s} = \frac{2}{\delta^2}\left(\frac{c_2 - \sqrt{t}Z_{1s}}{\sqrt{1-t}} + z_\beta\right)^2. \tag{12}$$

For the third step, we defined the conditional adaptive performance score (CAP) as follows:

$$\begin{aligned} &CAP(\delta) \\ &= \left\{\frac{\sum_{s=1}^{S} \left[n_{2fina\,l_s} - n_{2idea\,l_s}\right]_+ / n_{2idea\,l_s} + \left[\frac{2}{\delta^2}\left(\frac{c_2 - \sqrt{t}Z_{1s}}{\sqrt{1-t}} + z_{1-CP_s(\delta)}\right)^2 - n_{2idea\,l_s}\right]_- \Big/ \left[\frac{2}{\delta^2}\left(\frac{c_2 - \sqrt{t}Z_{1s}}{\sqrt{1-t}} + z_{0.5}\right)^2 - n_{2idea\,l_s}\right]_-}{S}\right\} \\ &\times 100 = \left\{\frac{\sum_{s=1}^{S} \left[n_{2fina\,l_s} - n_{2idea\,l_s}\right]_+ / n_{2idea\,l_s} + \left[n_{2fina\,l_s} - n_{2idea\,l_s}\right]_- \Big/ \left[\frac{2}{\delta^2}\left(\frac{c_2 - \sqrt{t}Z_{1s}}{\sqrt{1-t}} + z_{0.5}\right)^2 - n_{2idea\,l_s}\right]_-}{S}\right\} \times 100 \end{aligned} \tag{13}$$

where $CP_s(\delta)$ denotes $CP\big(Z_{1s}, t, n_{2final\_s};\ \delta_2 = \delta\big)$. Note that, considering n_{2ideal_s} and the reference under-power case of sample size $\frac{2}{\delta^2}\left(\frac{c_2-\sqrt{t}Z_{1s}}{\sqrt{1-t}} + z_{0.5}\right)^2$, which denotes a stage 2 sample size for which the conditional power equals 50% and can vary according to interim data Z_{1s}, the same value of $COS(\delta)$ or $CUP(\delta)$ can have different weights in $CAP(\delta)$. Therefore, if one calculates $CAP(\delta)$ to consider the utility of adaptive sample size designs in a trial simulation, we recommend simultaneously referring to $COS(\delta)$ and $CUP(\delta)$, which can be considered in the original scale of the sample size or probability.

Simulation Study

We conducted a simulation study in a simple setting to apply the adaptive performance measures considered in section of "A new adaptive performance measure" to various adaptive sample size designs with different required stage 2 sample size re-estimation criteria.

Sample Size re-estimation Criteria

(1) Delta-replacement criteria

Cui, Hung, and Wang (1999) proposed delta-replacement criteria as follows:

$$n_2^* = \left(\frac{\delta_{pre}}{\hat{\delta}_1}\right)^2 N_{initial} - n_1, \tag{14}$$

where n_2^* denotes the required stage 2 sample size re-estimated based on the interim result.

(2) Conditional power criteria

Proschan and Hunsberger (1995) proposed conditional power criteria that re-estimate the required n_2^* to meet $CP\big(Z_1, t, n_2^*;\ \delta_2 = \hat{\delta}_1\big) = 1 - \beta$, and it is calculated as follows.

$$n_2^* = \frac{2}{\hat{\delta}_1^{\,2}}\left(\frac{c_2 - \sqrt{t}Z_1}{\sqrt{1-t}} + z_\beta\right)^2, \tag{15}$$

where $CP\big(Z_1, t, n_2^*;\ \delta_2 = \hat{\delta}_1\big)$ is calculated under $\delta_2 = \hat{\delta}_1$ instead of $\delta_2 = \delta$, as in equation (10). Note that n_2^* can also be calculated in a usual fixed design sample size formula using the conditional error function $1 - \Phi\left(\frac{c_2-\sqrt{t}Z_1}{\sqrt{1-t}}\right)$ as a type I error level instead of α (Proschan and Hunsberger,1995). Bauer and Köenig (2006) pointed out that $CP\big(Z_1, t, n_2;\ \delta_2 = \hat{\delta}_1\big)$ was widely distributed and variable. Therefore, n_2^* estimated by conditional power criteria may be variable.

(3) Bayesian predictive power criteria with a non-informative prior

Spiegelhalter, Freedman, and Blackburn (1986) proposed a Bayesian predictive power approach for clinical trial monitoring as an alternative to the conditional power approach proposed by Halperin et al. (1982). Predictive power can be considered an unconditional prediction of conditional power that can take into account the uncertainty of the conditional power approach, which is assessed by two points of the hypothesis $\delta_2 = 0, \delta_{pre}$. Predictive power denoted by PP can be calculated using a weighted average of the conditional power function of δ_2, denoted by CP_{δ_2}, weighted by the posterior distribution of δ_2, denoted by $p(\delta_2|Z_1)$, as follows:

$$PP = \int CP_{\delta_2} p(\delta_2|Z_1)d\delta_2. \tag{16}$$

PP can be considered a conditional power based on the predictive distribution of stage 2. The predictive probability density function of $\hat{\delta}_2$, denoted by $f_p(\hat{\delta}_2|Z_1)$, is calculated as follows:

$$f_p(\hat{\delta}_2|Z_1) = \int f(\hat{\delta}_2;\ \delta_2)p(\delta_2|Z_1)d\delta_2, \tag{17}$$

where subscript p denotes the predictive distribution of stage 2 and statistics based on this distribution.

Expectation and variance of $f_p(\hat{\delta}_2|Z_1)$ are denoted by δ_p and σ_p^2, respectively, and conditional power is based on the predictive distribution of stage 2, denoted by CP_p, as follows:

$$CP_p = 1 - \Phi\left(\frac{c_2 - \sqrt{t}Z_1}{\sqrt{1-t}} - \frac{\delta_p}{\sigma_p}\right), \tag{18}$$

where Spiegelhalter, Freedman, and Blackburn (1986) show $CP_p = PP$.

Wang (2007) proposed predictive power criteria with non-informative prior criteria. Let the prior distribution of the effect size δ, denoted by $p(\delta)$, follow a normal distribution with prior mean and prior variance as follows:

$$p(\delta) \sim N(\delta_0, \sigma_0^2). \tag{19}$$

According to the Bayes rule, the posterior distribution of δ, given stage 1 data denoted by $p(\delta_2|Z_1)$, is calculated as follows:

$$\begin{aligned} p(\delta_2|Z_1) &= \frac{L(\delta_2|Z_1)p(\delta_2)d\delta_2}{\int L(\delta_2|Z_1)p(\delta_2)d\delta_2} \\ &\sim N\left(\frac{\frac{\delta_0}{\sigma_0^2} + \frac{n_1}{2}\hat{\delta}_1}{\frac{1}{\sigma_0^2} + \frac{n_1}{2}}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{n_1}{2}}\right). \end{aligned} \tag{20}$$

Considering non-informative priors with $\sigma_0^2 \sim \infty$, (20) reduces to the following:

$$p(\delta_2|Z_1) \sim N\left(\hat{\delta}_1, \frac{2}{n_1}\right). \tag{21}$$

According to $f(\hat{\delta}_2;\ \delta_2) \sim N\left(\delta_2, \frac{2}{n_2}\right)$, integral calculations such as equation (17) lead to a stage 2 predictive distribution of $f_p(\hat{\delta}_2|Z_1)$ as follows:

$$f_p(\hat{\delta}_2|Z_1) \sim N\left(\hat{\delta}_1, \frac{2}{n_1} + \frac{2}{n_2}\right). \tag{22}$$

According to equation (18), (22), required stage 2 sample size based on the Bayesian predictive power with a non-informative prior is calculated as
follows:

$$n_2^* = \left[\frac{1}{2}\left\{\hat{\delta}_1 \Big/ \left(\frac{c_2 - \sqrt{t}Z_1}{\sqrt{1-t}} + z_\beta\right)\right\}^2 - \frac{1}{n_1}\right]^{-1}. \tag{23}$$

(4) Bayesian predictive power criteria with an informative prior

All of the above criteria, including iii), do not depend on prior information and only use observed data in the clinical trial. One may think it should be avoided from using a Bayesian method with an informative prior in a confirmatory clinical trial. We propose using the prior in an adaptive sample size design, considering that a sample size determination based fully on prior information is very standard, and a sample size re-determination based fully on interim data can be risky. Note that we will still mainly assume the limited situation of a promising orphan drug, where some need for flexibility in initial and final sample sizes exists, as discussed in the introduction. Considering the situation of a promising effect of an orphan drug, one can target some prior range of clinically meaningful effect sizes denoted by PR as follows:

$$PR = [\delta_{lower}, \delta_{upper}], \tag{24}$$

where $\delta_{pre} \in PR$. Among PR, one may choose δ_{pre} to shorten the duration of a trial as much as possible without risking the loss of power. Now, we propose two types of prior distribution, denoted by $p(\delta) \sim N(\delta_0, \sigma_0^2)$, based on PR, for the Bayesian predictive power criteria. One type is as follows:

$$\delta_0 = \delta_{pre}, \sigma_0 = \frac{|\delta_{upper} - \delta_{lower}|}{2z_\alpha}, \tag{25}$$

where the width of the 95 percent coverage interval of the prior distribution equals the width of PR. Note that the smaller the interval the more informative a prior becomes. The second type is as follows:

$$\delta_0 = \delta_{pre}, \sigma_0 = |\hat{\delta}_1 - \delta_{pre}|. \tag{26}$$

If $\hat{\delta}_1$ is distant from δ_{pre}, the interim data may indicate an inconsistency between δ_{pre} and the true δ, and the weight of the prior distribution will be reduced. Similar calculations to equation (18) and (20) lead to the following two types of Bayesian predictive power criteria with an informative prior, based on the prior distributions denoted in equations (25), (26).

$$n_2^* = \left[\frac{1}{2}\left\{\frac{\delta_{pre}\Big/\left(\frac{|\delta_{upper} - \delta_{lower}|}{2z_\alpha}\right)^2 + \frac{n_1}{2}\hat{\delta}_1}{1\Big/\left(\frac{|\delta_{upper} - \delta_{lower}|}{2z_\alpha}\right)^2 + n_1/2} \Big/ \left(z_{A(Z_1)} + z_\beta\right)\right\}^2 - 1\Big/\left\{n_1 + 1/2\left(\frac{|\delta_{upper} - \delta_{lower}|}{2z_\alpha}\right)^2\right\}\right]^{-1} \tag{27}$$

$$n_2^* = \left[\frac{1}{2}\left\{\frac{\delta_{pre}/\left|\hat{\delta}_1 - \delta_{pre}\right|^2 + \frac{n_1}{2}\hat{\delta}_1}{1/\left|\hat{\delta}_1 - \delta_{pre}\right|^2 + n_1/2} \Big/ \left(z_{A(Z_1)} + z_\beta\right)\right\}^2 - 1/\left\{n_1 + 1/2\left|\hat{\delta}_1 - \delta_{pre}\right|^2\right\}\right]^{-1}. \tag{28}$$

Simulation Settings

We considered a placebo-controlled randomized two-arm clinical trial, as described in section of “A new adaptive performance measure”, and data will be simulated from the standard normal distribution. A prior range of clinically meaningful effect sizes is set as $PR = [0.2, 0.3]$, which corresponds to the required sample sizes of 392 and 174 per group, respectively. Corresponding to the above PR, δ_{pre} is set as $\delta_{pre} = 0.225, 0.275$, which corresponds to $N_{initial} = 310, 208$. Considering the situation of a promising orphan drug and a maximum sample size that may not be much larger than the initial setting, we set $\delta_{min} = 0.15$, corresponding to $N_{max} = 698$. Corresponding to $\delta_{upper} = 0.3$ of PR, we set $N_{min} = 174$. Note that $N_{min} = n_1$, with a scenario of $n_1 > 174$. The timing of the interim analysis is set in three patterns as $t = 0.25, 0.5, 0.75$ and c_1 and c_2 are set as O’Brien & Flemming-type boundaries, with $c_1 = 3.92$ and $c_2 = 1.96$, for $t = 0.25$; $c_1 = 2.78$ and $c_2 = 1.98$ for $t = 0.5$; and $c_1 = 2.33$ and $c_2 = 2.02$ for $t = 0.75$. Then, the final sample size is as follows:

$$N_{final} = \begin{cases} n_1, & if\ Z_1 \leq 0 \text{ or } Z_1 \geq c_1 \\ N_{min}, & if 0 < Z_1 < c_1 \text{ and } n_2^* < N_{min} - n_1 \\ n_1 + n_2^*, & if 0 < Z_1 < c_1 \text{ and } N_{min} - n_1 \leq n_2^* \leq N_{max} - n_1{}' \\ N_{max}, & if 0 < Z_1 < c_1 \text{ and } n_2^* > N_{max} - n_1 \end{cases} \tag{29}$$

where n_2^* is re-estimated based on the criteria shown above (1)–(4). We denoted each adaptive sample size design by $D_{replace}$, CP, $Noninfo$, $PP_{info}A$, and $PP_{info}B$, corresponding to equations (14), (15), (23), (27), and (28), respectively. Note that we set an important reference adaptive sample size design with conditonal power criteria under true δ instead of $\hat{\delta}_1$, denoted by $True$, to assess the adaptive performance of each design. We also set other reference adaptive sample size designs, which are group sequential designs with one interim analysis at t and are denoted by GSD_S, GSD_M, and GSD_L corresponding to $N_{GSD} = N_{initial}, (N_{initi\,al} + N_{max})/2, N_{max}$, respectively. Note that N_{GSD} denotes the maximum sample size of each group sequential designs, and N_{max} is the maximum sample size of this trial simulation as set above. According to $N_{initial} = 310, 208$ and $N_{max} = 698$, $N_{GSD} = 310, 504, 698$ and $208, 453, 698$, respectively.

Simulation Results

Figure 1 shows results comparing $POWER(\delta)$, $ASN(\delta)$, and $EAP(\delta)$ of each design. Two settings（$t = 0.5$, $\delta_{pre} = 0.225, 0.275$）are arrayed in two columns. The left column corresponds to $\delta_{pre} = 0.225; N_{inititial} = 310$, and the right column corresponds to $\delta_{pre} = 0.275; N_{initial} = 208$. Note that four settings（$t = 0.25, 0.75$, $\delta_{pre} = 0.225, 0.275$）are omitted from the figures as important trends are similar to the two above. Of the three reference lines on the horizontal axis, the left denotes the lower boundary of PR ($\delta_{lower} = 0.2$), the right denotes the upper boundary ($\delta_{upper} = 0.3$), and the middle denotes $\delta_{pre} = 0.225, 0.275$. The reference line for $POWER(\delta)$ denotes a nominal power of $1 - \beta = 0.8$. First, we compared $POWER(\delta)$ between the group sequential designs (GSDs) and sample size re-estimation designs

(SSRs). GSD_M, GSD_L had higher $POWER(\delta)$ than SSRs in every setting and reached nominal power at all effect sizes over PR. On the other hand, the same sample size of GSD_S as the initial sample size of SSRs showed consistently lower $POWER(\delta)$ than SSRs and was much lower than nominal at many effect sizes in PR, especially in the right column. Only $Noninfo$ among the SSRs showed nominal power in every setting over all PR. Then, we compared $POWER(\delta)$ among the SSRs. $D_{replace}$ and CP showed similar $POWER(\delta)$ and mostly had nominal power, except for $t = 0.25, \delta_{pre} = 0.275$. The proposed $PP_{info}A$ and $PP_{info}B$ showed the lowest $POWER(\delta)$ among the SSRs. $PP_{info}B$, for which the prior distribution had weight based on interim data, showed a trend of higher $POWER(\delta)$ than that of $PP_{info}A$, with a prior having a constant weight according to the width of PR. Note that unless the timing of the interim analysis is not early ($t = 0.25$), $PP_{info}B$ was not less than 5%, compared to the nominal power in most effect sizes in PR even with small initial sample size ($\delta_{pre} = 0.275$). The important reference SSR is the ideal SSR denoted by $True$ and re-estimates required stage 2 sample size based on every true effect size under the limitation of the initial and maximum sample sizes. $True$ can be considered a robust design with less diversity of $POWER(\delta)$ with different effect sizes over PR. The proposed $PP_{info}A$ and $PP_{info}B$ showed a trend of lower power, but smaller differences from $True$ with every setting. Secondly, we compared $ASN(\delta)$ between each design. The upper of the two reference lines on the vertical axis showed an initial sample size of $N_{initial} = 392$, which has just the nominal power for the lower bound of PR ($\delta_{lower} = 0.2$) with a fixed design. The lower reference line shows $N_{min} = 174$, corresponding to the upper bound ($\delta_{upper} = 0.3$). Compared to $POWER(\delta)$, $ASN(\delta)$ trended larger withmore powerful designs. GSD_M, GSD_L, and $Noninfo$ were beyond the reference sample size of $N_{initial} = 392$ and considered to be over-sized at every effect size in PR. The proposed $PP_{info}A$ and $PP_{info}B$ showed smaller $ASN(\delta)$ and small difference from $True$ with every setting. Thirdly, we compared $EAP(\delta)$, which integrates the expected under-power and over-size in the long run, between designs. GSD_M, GSD_L and $Noninfo$ showed larger $EAP(\delta)$ in PR, indicating over-size in PR, even considering the under-power of other designs. GSD_S also showed larger $EAP(\delta)$ in PR due to under-power. The proposed $PP_{info}A$ and $PP_{info}B$ had small differences from $True$ with every setting. $PP_{info}A$ had the minimum $EAP(\delta)$ around δ_{pre}, and change in $EAP(\delta)$ was moderate at δ, different from δ_{pre}, while GSD_S, GSD_M, GSD_L had a large slope of $EAP(\delta)$.

Figure 2 shows results comparing $CAP(\delta), CUP(\delta)$, and $COS(\delta)$ between the designs. $CAP(\delta)$ not only integrates under-power and over-size conditional on an interim result but also accommodates variability in the adaptive performance. Considering variability, differences in the adaptive performance of GSDs and SSRs became clear. The proposed $PP_{info}A$ and $PP_{info}B$ showed a trend of smaller $CAP(\delta)$, especially in the left column with $N_{initial} = 310$, than $CP, D_{replace}, Noninfo$. Next, we compared $CUP(\delta)$ and $COS(\delta)$, which can be assessed on the original scale, considering the variability in the adaptive performance, and, as we recommend in the last of section of “A new adaptive performance measure”, can be referred to simultaneously with $CAP(\delta)$. $CUP(\delta)$ indicates the average loss of conditional power in each design compared to the nominal value of $1 - \beta$ in the probability scale. Note that $CUP(\delta)$ only counts as a loss, but does not count as excess and can take into account variability in the conditional power. $PP_{info}A$ and $PP_{info}B$, which showed a trend of lower $POWER(\delta)$, had no more than 10 percent of $CUP(\delta)$for most of the effect sizes in PR, which contrasts with GSD_S, using the same initial sample size. For the left column with a greater initial sample size, $PP_{info}A$ and $PP_{info}B$ had no more than 5 percent of $CUP(\delta)$ in PR and good performance compared to the large $ASN(\delta)$ of GSD_M, GSD_L and $Noninfo$. $COS(\delta)$ indicates the average over-size of the final stage 2 sample size $(N_{final} - n_1)$ of each design compared to the exact size with a conditional power of $1-\beta$ on the scale of the sample size. Note that $COS(\delta)$ only counts over-size but does not count under-size and can consider variability in the final stage 2 sample size. The reference line on the vertical axis indicates the maximum over-size of the fixed design with $N_{initial} = 392$ based on $\delta = 0.2$ for $\delta = 0.3$. $PP_{info}\,A$ showed a small $COS(\delta)$, especially with $\delta_{pre} = 0.275$, $N_{initial} = 208$, assuming likely settings for a promising orphan drug, in which one can expect efficacy but a longer entry period and may justify minimizing the final sample size to address unmet medical need immediately as possible.

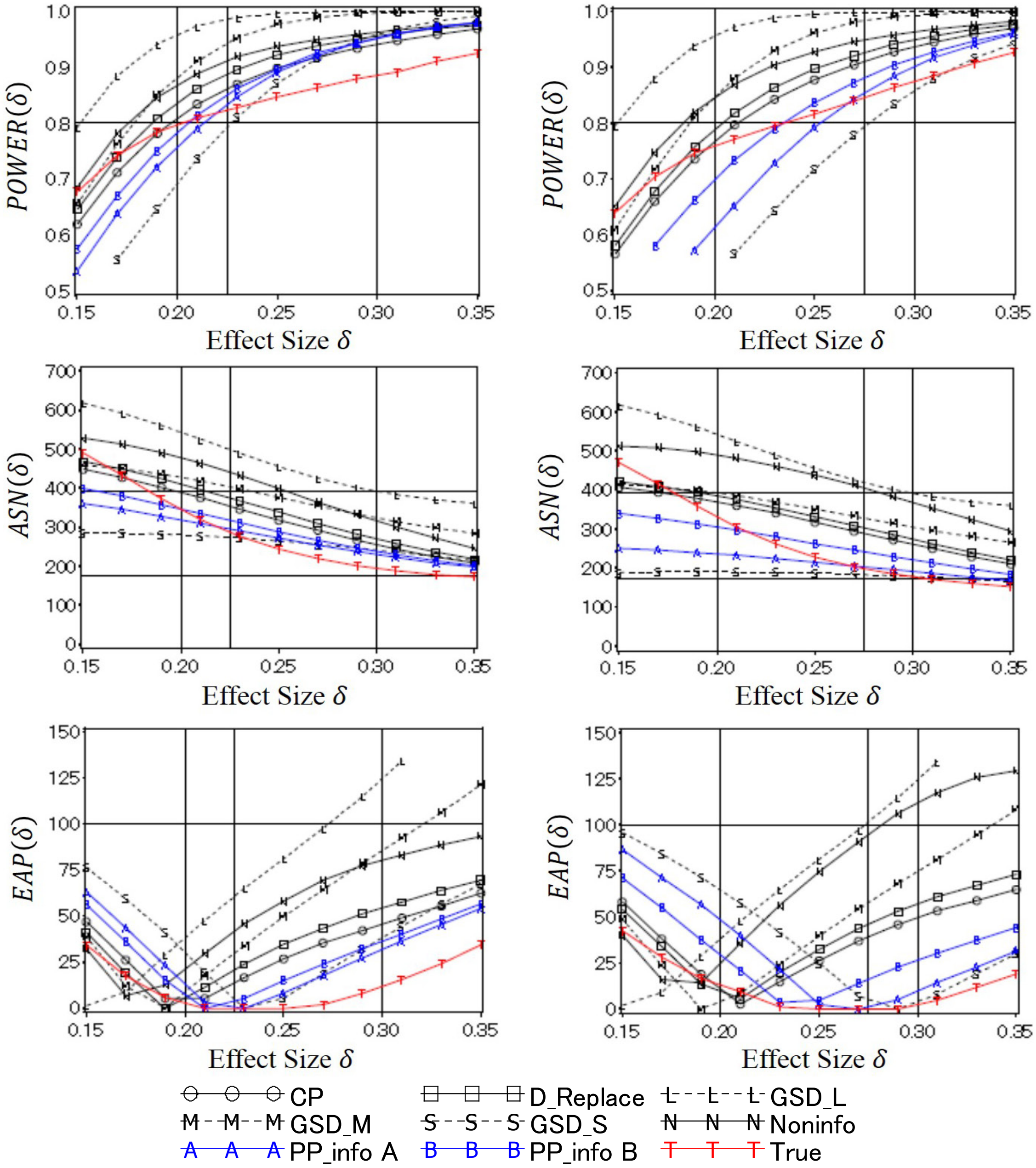

Figure 1. Results of a simulation study comparing $POWER(\delta)$, $ASN(\delta)$, $EAP(\delta)$ of each adaptive sample size design with the interim analysis at $t = 0.5$.

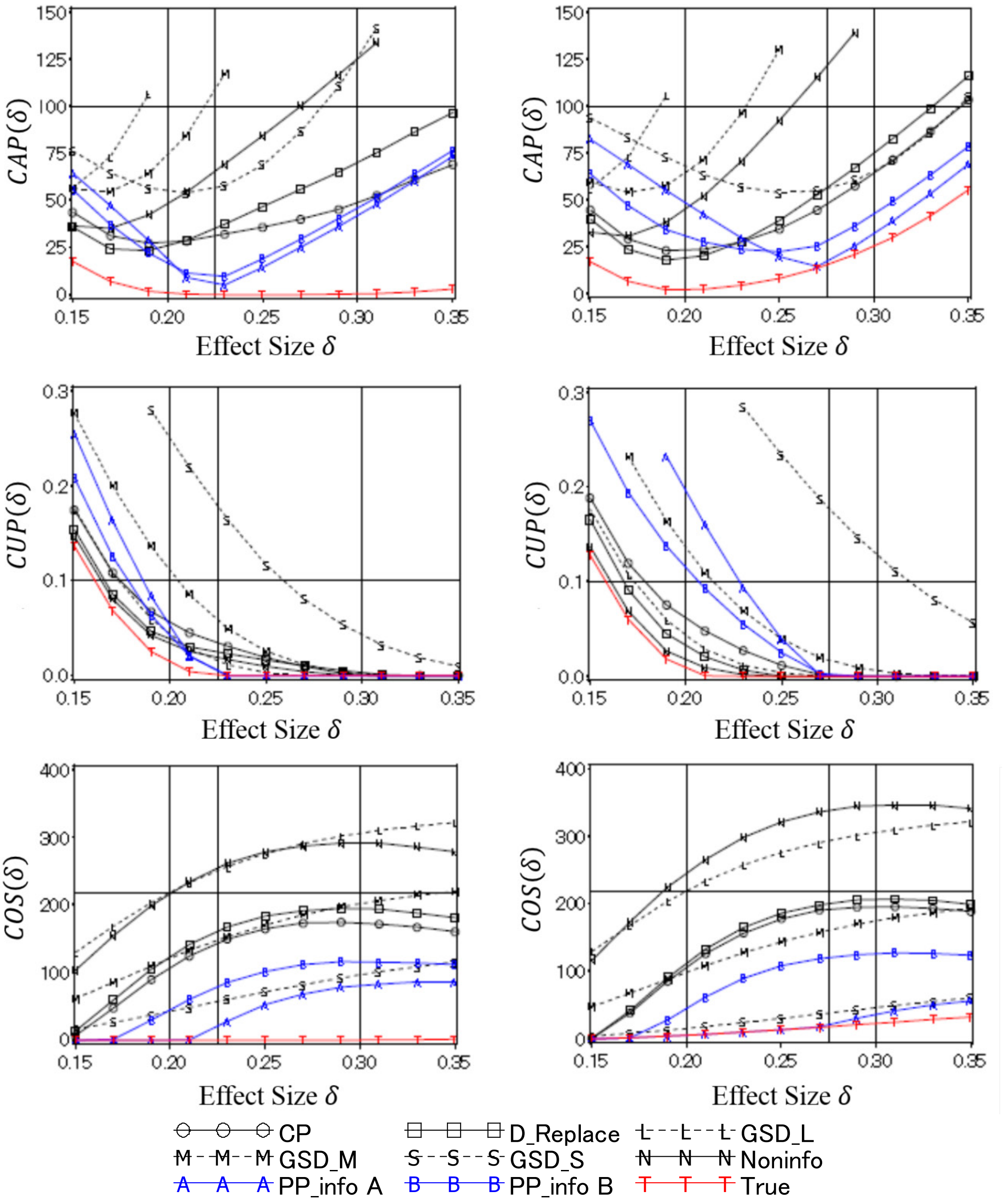

Figure 2. Results of simulation study comparing $AP(\delta)$, $CUP(\delta)$, $COS(\delta)$ of each adaptive sample size designs with interim analysis at $t = 0.5$.

A review Example of SSR

Here, we will present a review example of SSR. Although it is not a situation of promising orphan drug development, it may be one of the rare SSR examples in which new drugs are approved in Japan. Tiotropium, which is mainly used in patients with chronic obstructive pulmonary disease in Japan, has recently been approved for asthma patients. In pivotal twin trials comparing tiotropium 5 μg and placebo for severe asthma patients with three co-primary endpoints, sample size re-estimation was planned and conducted for the third co-primary endpoint (time to first severe asthma exacerbation evaluated in 48 weeks), based on pooled trial data. The first two co-primary endpoints (change from baseline of peak and trough FEV_1(L) at 24 weeks) required 150 subjects per group for more than 90 percent power. The sample size estimation of the third endpoint described in the protocol is presented in Table 1. According to the first table of Table 1, which shows the distribution of the number of severe asthma exacerbations estimated from an omalizumab trial, the proportion of patients with no exacerbations are assumed to be about 0.7 to 0.8. Note that omalizumab is a biological drug that can be used for severe asthma patients in Japan.

Under the omalizumab assumption, the prior hazard ratio was set to 0.562, as shown at the middle row in the second table of Table 1, which indicates the sample size sensitivity for the third co-primary endpoint. The required sample size as the pooled total is estimated to be 574. If an interim estimate of the hazard ratio exceeds 0.6 based on 65 events, the pooled total sample size should be increased to 812 according to the table. This corresponds to an increase in target hazard ratio from 0.562 to 0.611, which may require only a minor change in the targeted treatment effect. The final p-value is computed by the weighted Z statistic proposed by Cui, Hung, and Wang (1999). In fact, a pre-specified interim analysis was performed once at 65 events for the pooled total; as a result, the sample size was increased to approximately 400 patients per trial from 300 that was set at the trial start. Results were cited in the PMDA review report.

Table 1
Distribution of the number of severe asthma exacerbations estimated from an omalizumab trial and sample size sensitivity for the third co-primary endpoint, time to first severe asthma exacerbation in the48th week, based on pooled tiotropium trial data. These tables are reproduced from the protocol (http://www.nejm.org/doi/suppl/10.1056/NEJMoa1208606/suppl_file/nejmoa1208606_protocol.pdf)

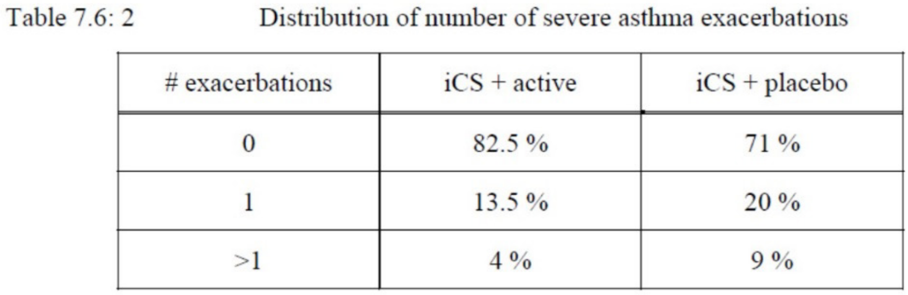

Table 7.6: 2 Distribution of number of severe asthma exacerbations

# exacerbations	iCS + active	iCS + placebo
0	82.5 %	71 %
1	13.5 %	20 %
>1	4 %	9 %

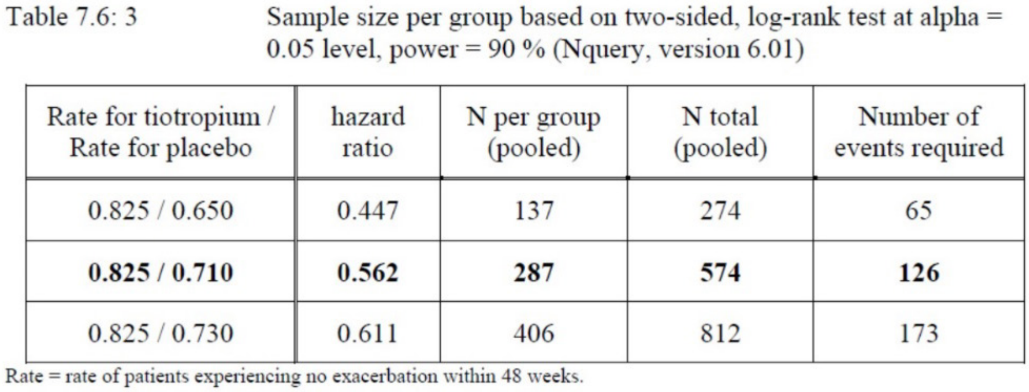

Table 7.6: 3 Sample size per group based on two-sided, log-rank test at alpha = 0.05 level, power = 90 % (Nquery, version 6.01)

Rate for tiotropium / Rate for placebo	hazard ratio	N per group (pooled)	N total (pooled)	Number of events required
0.825 / 0.650	0.447	137	274	65
0.825 / 0.710	**0.562**	**287**	**574**	**126**
0.825 / 0.730	0.611	406	812	173

Rate = rate of patients experiencing no exacerbation within 48 weeks.

In terms of the first two co-primary endpoints, the superiority to placebo was demonstrated. The differences, 95% confidence intervals and p-values between tiotropium and placebo in terms of change from baseline of peak FEV_1(L) at 24 weeks for each trial were 0.086 [0.020, 0.152], p=0.0110, 0.154 [0.091, 0.217], p<0.0001, respectively. The differences, 95% confidence intervals and p-values between tiotropium and placebo in terms of change from baseline of trough FEV_1(L) at 24 weeks for each trial were 0.088 [0.027, 0.149], p=0.0050, 0.111 [0.053, 0.169], p=0.0002. In the PMDA review, trough FEV_1 is considered one of the most important endpoints to evaluate a controller drug with a long-acting bronchodilator such as tiotropium. Therefore, the results of trough $FEV_{1,}$ which was consistent in pivotal twin trials, were important to the assessment of asthma indication. The primary result of the third endpoint was as follows: the adjusted hazard ratio, 95% confidence interval and the adjusted p-value (Cui, Hung and Wang, 1999) between tiotropium and placebo in terms of time to first severe asthma exacerbation evaluated in 48 weeks was 0.77 [0.60, 0.98], p=0.0343. Statistical review points relating to SSR are as follows: Accrual curves, demographic and baseline disease characteristics, and the first two co-primary endpoints were compared between the before and after interim analyses in each trial. Similar comparisons were also considered for pooled endpoints, not only the third primary but also the secondary exacerbation endpoint and symptomatic asthma exacerbation. The PMDA discussed whether the third primary result should be included in the section of trial results of the label and concluded that it should be included.

Discussion

In this paper, we proposed new adaptive performance assessment measures to evaluate the utility of an adaptive sample size design in a trial simulation, especially in the situation of a promising orphan drug, with a long accrual period and the expectation of benefit and low risk. As proposed, $CAP(\delta)$, $CUP(\delta)$, and $COS(\delta)$ take into account the mean error of an adaptive decision, given various results in the interim analysis, and can measure different performance from $POWER(\delta)$, $ASN(\delta)$, $EAP(\delta)$, which measure the overall trial performance in the long run. We believe that the overall trial performance strongly depends on the initial settings of design parameters, prior information, and initial careful planning, considering $POWER(\delta)$, required sample size, accrual duration, feasibility, and other factors. Given initial careful planning, a decision to change from the initial design setting based on the results of the interim analysis should be made conservatively. Therefore, sample size re-estimation may be an optional aid to ease limitations in the initial planning. The proposed Bayesian sample size re-estimation criteria may fit to the above purpose of SSR. Our brief example of a trial simulation showed that the Bayesian criteria of $PP_{info}A$ and $PP_{info}B$ could approach $True$, which is the hypothetical adaptive sample size design given the true effect size, especially near the initial setting of δ_{pre}, denoted by PR. On the other hand, our simulation showed that no design could necessarily cope with a range of effect sizes in an efficient and precise manner. In addition, considering the low probability of efficacy if a trial is stopped after the interim analysis and the high probability of reaching N_{max}, GSD without SSR showed a limited ability to cope with the potential gap between δ_{pre} and the true effect size and variability in interim results. Therefore, it is important to balance the potential risk and utility of each design, based on prior belief or expectations for the risk/benefit profile of an investigational drug. The review example of SSR provided in section of "A review example of SSR", though no large issues were presented, showed that to gain highly conclusive evidence for the third endpoint, the target effect size and/or range of sample size sensitivity should be set conservatively (wider), considering the effect size observed in a similar trial of an inhaled combination of corticosteroid and long-acting bronchodilator drugs. It is better to conduct another larger-sized confirmatory trial to confirm the reproducibility of the reduction in severe exacerbation of severe persistent asthma considering the moderate effect size.

References

Bauer, P. Köhne, K. (1994) Evaluation of experiments with adaptive interim analyses. *Biometrics*,50:1029-1041.

Bauer, P. König, F. (2006) The reassessment of trial perspectives from interim data－a critical view. *Statistics in Medicine*, 25: 23-36.

Brannath, W.; König, F. and Bauer, P. (2006) Estimation in flexible two stage designs. *Statistics in Medicine*, 25: 3366-3381.

Cui, L.; Hung, H.M.J. and Wang, S. (1999) Modification of sample size in group sequential clinical trials. *Biometrics*, 55: 853-857.

FDA. Adaptive design clinical trials for drugs and biologics. Guidance for industry, 2010.

FDA. Briefing document in review of Sunitinib (PNET), April 12, 2011, ODAC.Available from:

http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/OncologicDrugsAdvisoryCommittee/UCM251683.pdf (accessed 24 October 2016)

Halperin, M.; Lan, K.K.G.; Ware, J.H.; Johnson, N.J. and DeMets, D.L. (1982) An aid to data monitoring in long-term clinicaltrials. *Controlled Clinical Trials*, 3: 311-323.

Hung, H.J.; Cui, L.; Wang, S.J. and Lowrence, J. (2005) Adaptive statistical analysis following sample size modification based on interim review of effect size. *Journal of Biopharmaceutical Statistics*, 15: 693-706.

Jennison, C.; Turnbull, B.W. (2003) Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, 22: 971-993.

Kerstjens H.A.M., Engel M., Dahl R., et al. (2012) Tiotropium in asthma poorly controlled with standard combination therapy. *New England Journal of Medicine*, 367:1198-207. Also the trial protocol is available from: http://www.nejm.org/doi/suppl/10.1056/NEJMoa1208606/suppl_file/nejmoa1208606_protocol.pdf (accessed 24 October 2016)

Lan, K.K.G.; Simon, R. and Halperin, M. (1982) Stochastically curtailed tests in long-term clinical trials. *Sequential Analysis*, 1: 207-219.

Liu, Q.; Proschan, A.A. and Pledger, G.W. (2002) A unified theory of two-stage adaptive designs. *Journal of the American Association*, 97: 1034-1041.

Liu, D.F. and Cui, L. (2008) Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. *Statistics in Medicine*, 27: 584-596.

Pharmaceuticals and Medical Devices Agency. (2014) ReviewReport of Tiotropium. Pharmaceuticals andMedical Devices Agency, 10 October 2016 (in Japanese). Available from:
http://www.pmda.go.jp/drugs/2014/P201400147/530353000_22200AMX00227_A100_2.pdf (accessed 24 October 2016).

Proschan. M. and Hunsberger, S. (1995) Designed extension studies based on conditional power. *Biometrics*, 51: 1315-1324.

Spiegelhalter, D.J.; Freedman, L.S. and Blackburn, P.R. (1986) Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials*; 7: 8-17.

Tsiatis, A.A. and Mehta, C. (2003) On the efficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90:367-378.

Wang, M.D. (2007) Sample size reestimation by Bayesian prediction. *Biometrical Journal*, 49: 365-377.