Robustness and corrections for sample size adaptation strategies based on effect size estimation

Daniele de Martini

Robustness and corrections for sample size adaptation strategies based on effect size estimation

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2010-0247.R2
Manuscript Type:	Original Paper
Date Submitted by the Author:	20-Feb-2011
Complete List of Authors:	De Martini, Daniele; Università di Milano - Bicocca, Dipartimento DIMEQUANT
Keywords:	structural bias, conservativeness, sample size estimation
Abstract:	The robustness of the adaptation of the sample size for a phase III trial based on phase II data is studied -- when phase III is lower than phase II effect size. Conservative sample size estimation strategies are compared. When the rate between phase III and phase II effect size is greater than 0.8, Calibrated Optimal Strategy provides acceptable results. A correction for balancing the structural bias is introduced, based on a postulation of the bias. When the postulated correction is right, or even smaller than necessary, COS works well. A higher than necessary correction should be avoided.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
cos_sse_diff_scen3.tex	

Robustness and corrections for sample size adaptation strategies based on effect size estimation

Daniele De Martini
Dipartimento DIMEQUANT - Università degli Studi di Milano - Bicocca
Via Bicocca degli Arcimboldi 8, 20126 Milano - Italia
E-mail: daniele.demartini@unimib.it

Summary A study on the robustness of the adaptation of the sample size for a phase III trial on the basis of existing phase II data is presented – when phase III is lower than phase II effect size. A criterion of clinical relevance for phase II results is applied in order to launch phase III, where data from phase II cannot be included in statistical analysis. The adaptation consists in adopting the conservative approach to sample size estimation, which takes into account the variability of phase II data. Some conservative sample size estimation strategies, Bayesian and frequentist, are compared with the calibrated optimal γ conservative strategy (viz. COS) which is the best performer when phase II and phase III effect sizes are equal. The Overall Power (OP) of these strategies and the mean square error (MSE) of their sample size estimators are computed under different Scenarios, in the presence of the structural bias due to lower phase III effect size, for evaluating the robustness of the strategies. When the structural bias is quite small (i.e. the ratio of phase III to phase II effect size is greater than 0.8), and when some operating conditions for applying sample size estimation hold, COS can still provide acceptable results for planning phase III trials, even if in bias absence the OP was higher.

Main results concern the introduction of a correction, which affects just sample size estimates and not launch probabilities, for balancing the structural bias. In particular, the correction is based on a postulation of the structural bias; hence it is more intuitive and easier to use than those based on the modification of type I or/and type II errors. A comparison of corrected conservative sample size estimation strategies is performed in the presence of a quite small bias. When the postulated correction is right, COS provides good OP and the lowest MSE. Moreover, the OPs of COS are even higher than those observed without bias, thanks to higher launch probability and a similar estimation performance. The structural bias can therefore be exploited for improving sample size estimation performances. When the postulated correction is smaller than necessary, COS is still the best performer, and it also works well. A higher than necessary correction should be avoided.

Keywords: structural bias; conservativeness; launch threshold; Overall Power; postulated correction; sample size estimation.

1 Introduction

Sample size estimation (SSE) consists in estimating the sample size required by a statistical test for achieving a predefined power. In the parametric framework the power is a function of the true effect size (ES), so that SSE can be accomplished through the estimation of this ES. In practice, the true ES is unknown (being this latter the object of the research) and it can be estimated on the basis of data from a pilot study. Subsequently, the sample size for the further study is estimated. For simplicity, and to establish SSE in the context of clinical trials, we assume that the pilot and the subsequent experiments are a phase II and a phase III trial, respectively. SSE is an intuitive practice that has been adopted by many authoritative authors (see for example Rosner, 2005, Ch.8; Efron and Tibshirani, 1993, Ch.25).

In the last decade, the conservative approach to SSE (i.e. CSSE), which accounts for the variability of phase II data, has been proposed under both the frequentist (Shao and Chow, 2002, Sec.5.2) and the Bayesian approach (Chuang-Stein, 2006). The introduction, according to Wang et al.(2006), of a launch threshold in the setting of CSSE (i.e. phase III is launched, and its sample size is estimated, only when phase II shows effect size over a predefined threshold of a certain clinical importance) represents a strong link between theory and practice: on one hand the launching threshold rationalizes a step often adopted in practice, and on the other hand it induces an upper bound for sample size estimators, which is essential for practical purposes.

Some papers report comparisons among different CSSE strategies. Wang et al. (2006) evaluated three frequentist CSSE strategies, mainly on the basis of launch probability and of the Average Power of phase III (i.e. the average of the random power given by the randomness of sample size estimators, **provided that the phase III trial was launched**). These performance indicators were computed under four different Scenarios: in Scenario 1 the effect size of phase II was equal to that of phase III; in Scenarios 2-4 phase III effect size was lower than that of phase II. The authors concluded by suggesting the adoption of the conservative strategy consisting in the use of the effect size observed in phase II minus one time its standard error, in order to then compute the sample size for phase III. Note that in Scenarios 2-4 the authors studied the robustness of the three strategies.

Let us now introduce the Overall Power (OP) of the phase II *and* the phase III, that is the probability to launch times the Average Power. Fay et al. (2007) compared, mainly on the basis of a modified for continuity version of the OP, the simple pointwise frequentist approach with two Bayesian techniques. Scenario 1 alone was considered and the authors suggested the adoption of a corrected Bayesian strategy. **(In detail, to compute the OP Fay et al. (2007) set the launch threshold at $-\infty$ and when the effect size was estimated to be under the null hypothesis they estimated the sample size for the phase III, for continuity, to be ∞; then, in these cases, and when the alternative hypothesis was true, i.e. when estimating the sample size has a practical sense, the phase III power was considered to be 1. Further details and comments can be found in De Martini, 2010.)**

It is worth noting that, from the perspective of statistical methodology classification, Hung et al.(2006) argued that SSE, and CSSE too, can be viewed as a kind of adaptation by design.

Note also that the OP evaluates the performance of SSE strategies on phase II and phase III (i.e. independently on phase II results) and it is therefore indicated for comparing strategies globally, where the Average Power becomes of interest once the phase II succeeded, say at the beginning of phase III.

Recently, we compared (De Martini, 2010) some frequentist and some Bayesian strategies with a new conservative strategy based on the calibration of the optimal amount of conservativeness - COS. To evaluate the results the OP of the different strategies, as well as the mean and the MSE of sample size estimators, were computed. The launch threshold was adopted. Bayesian strategies performed poorly since they showed a very high mean and/or MSE of sample size estimators. COS clearly performed better than the other frequentist conservative strategies, both in terms of OP and of sample size estimators behavior **(viz. MSE)**. Costs and experimental times are, therefore, considerably reduced and standardized. These results were computed solely under Scenario 1.

Broader and more heterogeneous patient populations are often pursued in phase III clinical trials as compared to in phase II studies. This might induce larger variability and/or lower average differences between drugs in primary endpoints variables. Both phenomena imply that standardized phase III effect sizes are lower than phase II ones. Wang et al.(2006) modeled these situations in Scenarios 2-4.

Our aim is now to study the robustness of some frequentist, Bayesian and optimized (viz. COS) CSSE strategies through the evaluation of their performances under these latter Scenarios. Then, some techniques for correcting the difference between phase III and phase II effect size, that are usually adopted under these Scenarios, are discussed and a new one is introduced. The robustness of CSSE strategies improved by this correction technique are, hence, examined.

The paper is as follows: Section 2 deals with the theoretical framework, and Section 3 with CSSE strategies which will be compared, under Scenarios 2-4, in Section 4. Section 5 is devoted to the techniques of bias correction and to the application of the new one to CSSE strategies. In Section 6, a comparison among the corrected versions of our CSSE strategies is presented; Section 7 contains the discussion, and Section 8 the conclusions.

2 Theoretical framework

In accordance with Wang et al. (2006), we focus on the one-sided alternatives test in a comparison of the means of two normal distributions in phase III, as we did De Martini (2010). The common variance σ^2 is assumed to be known and equal to 1, in phase III as well as in phase II. The statistical hypotheses are, therefore,

$H_0 : \mu_{3,T} = \mu_{3,C}$ vs $H_1 : \mu_{3,T} > \mu_{3,C}$, and the standardized phase III effect size is $\delta_3 = (\mu_{3,T} - \mu_{3,C})/\sigma = \mu_{3,T} - \mu_{3,C}$.

Being $m_1 = m_2 = m$ the sizes of the samples drawn from each group in phase III, the test statistic is $T_m = \sqrt{m/2}(\bar{X}_m - \bar{Y}_m)$. Given the type I error α, the true power of the phase III trial is $\pi_{\delta_3}(m) = P_{\delta_3}(T_m > z_{1-\alpha}) = \Phi(\delta_3\sqrt{m/2} - z_{1-\alpha})$, where $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$, and Φ is the distribution function of the standard normal.

The ideal sample size per group for phase III is, therefore, $M_I = \min\{m \,|\, \pi_{\delta_3}(m) > 1-\beta\} = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/\delta_3^2 \rfloor + 1$, where $1-\beta$ is the power to be achieved. In practice we are required to estimate M_I, since δ_3 is unknown.

The effect size of phase II is $\delta_2 = \mu_{2,T} - \mu_{2,C}$. Being $n_1 = n_2 = n$ the sample size for each group in phase II, let $d^{\bullet}_{2,n}$ be a generic estimator of δ_3 based on phase II data, where "$\bullet$" represents a generic CSSE strategy. We apply the launching criterion introduced by Wang et al.(2006): δ_{0L} represents the launch threshold and phase III is launched on condition that $d^{\bullet}_{2,n} > \delta_{0L}$. In this case, M_I is estimated by $M^{\bullet}_n$, i.e. the sample size estimator of $\bullet$, usually based on $d^{\bullet}_{2,n}$. Note that $M^{\bullet}_n$ is a discrete random variable whose distribution depends on δ_2. Finally, assume that there exists a maximum for $M^{\bullet}_n$ and that this maximum corresponds to $M_{\max} = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/\delta_{0L}^2 \rfloor + 1$. For example, being $\bullet$ the fixed-γ conservative strategy, we have that $d^{\bullet}_{2,n} = d^{\gamma}_{2,n} = d_{2,n} - z_{\gamma}/\sqrt{n/2}$, where $d_{2,n} = \bar{X}_n - \bar{Y}_n$ is the pointwise estimator of δ_3, and $M^{\gamma}_n = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/(d^{\gamma}_{2,n})^2 \rfloor + 1$.

Hence, the OP (i.e. the probability of rejecting during phase III, when M_I is estimated on the basis of phase II data) is:

$$OP_n(\bullet) = \sum_{m=2}^{M_{\max}} P_{\delta_2}(M^{\bullet}_n = m)\pi_{\delta_3}(m) \tag{1}$$

The average of sample size estimators is then $E[M^{\bullet}_n | d^{\bullet}_{2,n} > \delta_{0L}]$, and their mean square error (MSE) is $E[(M^{\bullet}_n - M_I)^2 | d^{\bullet}_{2,n} > \delta_{0L}]$.

Remark 1. Although we used δ_2 and δ_3 for defining the effect sizes of phase II and phase III, it should be noted that the effect the drug has is not a function of the development phase of the sponsor. Instead, there are different treatment effects for the population studied for particular phase II and phase III protocols. Nevertheless, when $\delta_2 \neq \delta_3$ a structural bias arises within SSE.

3 CSSE strategies

In this Section five different CSSE strategies are recalled and are to be compared in Section 4 in order to evaluate their robustness. For each strategy, motivation for inclusion in our study is explained.

Pointwise strategy (PWS). It is the fixed-γ conservative strategy with $\gamma = 50\%$: in practice, the observed effect size $d_{2,n}$ is adopted and the sample size estimator is $M_n^{50\%} = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/(d_{2,n})^2 \rfloor + 1 = M_n$. This strategy is considered because it is the simplest, and also because the results regarding the estimation of the sample size under Scenario 1 were quite good (De Martini, 2010).

One standard error conservative strategy (1SES). In this strategy, 1 standard error is subtracted from $d_{2,n}$ to obtain a conservative estimate of δ_3. This corresponds to 84.1% conservative strategy, so that $d_{2,n}^{\bullet} = d_{2,n}^{84.1\%}$ and M_I is estimated by $M_n^{84.1\%}$. Although 1SES did not provide good results under Scenario 1 (De Martini, 2010), it is included in this comparison because Wang et al.(2006), who also considered Scenarios 2-4, deem it should be adopted.

Third quartile conservative strategy (3QS). To adopt 75% conservative strategy appears a reasonable choice, because in this way the probability of planning an underpowered experiment is reduced to the fixed rate of $1/4$. Hence, $d_{2,n}^{\bullet} = d_{2,n}^{75\%}$ and the sample size estimator is $M_n^{75\%}$. We consider this fixed-γ strategy too because the performances of 3QS in terms of OP under Scenario 1 were closer to COS than those of PWS and 1SES.

Calibrated optimal strategy (COS). This strategy is based on the fact that if a fixed-γ conservative strategy is adopted, then its OP (say $OP_n(\gamma)$) turns out to be a concave function of γ. Consequently, we introduced an optimization for $OP_n(\gamma)$ (De Martini, 2010), with the constraint of not exceeding $1-\beta$. Being that optimum (say $\gamma_{O,n}$) unknown, it can be estimated through the plug-in principle and its estimate is $g_{O,n}$. With COS $d_{2,n}^{\bullet} = d_{2,n}$ and the sample size estimator is $M_n^{g_{O,n}} = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/(d_n^{g_{O,n}})^2 \rfloor + 1$. COS performed very well under Scenario 1 (De Martini, 2010), but it has not yet been studied under different Scenarios.

Bayesian truncated strategy (BAT). In a Bayesian framework, the posterior distribution of δ_2 given $d_{2,n}$ is, with noninformative prior, $N(d_{2,n}, 2/n)$, and it is used for making inference on δ_3. Consequently the Bayesian estimate of the true power is $\pi_{Ba_n}(m) = \int_{-\infty}^{\infty} \pi_z(m)\phi_{d_{2,n},2/n}(z)\,dz$. The simple Bayesian estimate of the sample size, suggested by Chuang-Stein (2006) and Fay et al.(2007), is, therefore, $M_n^{BaS} = \min\{m \,|\, \pi_{Ba_n}(m) > 1-\beta\}$. Note that in some circumstances $\lim_{m\to\infty} \pi_{Ba_n}(m) < 1-\beta$, so that M_n^{BaS} does not exist, and in other circumstances M_n^{BaS} can be simply higher than $M_{\max}$. Consequently, we defined (De Martini, 2010) the truncated Bayesian sample size estimator $M_n^{BaT} = \min\{M_n^{BaS}, M_{\max}\}$. With BAT, $d_{2,n}^{\bullet} = d_{2,n}$. BAT strategy is included in this comparison because under Scenario 1 it was the best Bayesian performer, although its sample size estimator showed very high average and/or MSE (see De Martini, 2010).

Remark 2. With γ conservative strategies the OP in (1) simplifies to

$$OP_n(\gamma) = \sum_{m=2}^{M_{\max}} \{\Phi(\sqrt{\frac{n}{2}}(\frac{z_{1-\alpha}+z_{1-\beta}}{\sqrt{(m-1)/2}} - \delta_2) + z_\gamma) - \Phi(\sqrt{\frac{n}{2}}(\frac{z_{1-\alpha}+z_{1-\beta}}{\sqrt{m/2}} - \delta_2) + z_\gamma)\}\Phi(\sqrt{\frac{m}{2}}\delta_3 - z_{1-\alpha}) \tag{2}$$

Without loss of generality, we assume $\delta_2 = \delta_3/k$ so that (2) becomes:

$$OP_n(\gamma) = \sum_{m=2}^{M_{\max}} \{\Phi(\sqrt{\frac{n}{2}}(\frac{z_{1-\alpha}+z_{1-\beta}}{k\sqrt{(m-1)/2}} - \delta_2) + z_\gamma) - \Phi(\sqrt{\frac{n}{2}}(\frac{z_{1-\alpha}+z_{1-\beta}}{k\sqrt{m/2}} - \delta_2) + z_\gamma)\}\Phi(\sqrt{\frac{m}{2}}k\delta_2 - z_{1-\alpha}) \tag{3}$$

Remark 3. It should be noted that Fay et al.(2007) considered the variance to be unknown and that they used the fiducial distribution of the effect size, actually not a Bayesian posterior. Nevertheless, it is easy to show that in case the variance is known the fiducial distribution reduces to the Bayesian posterior with noninformative prior, i.e. $N(d_{2,n}, 2/n)$.

4 Evaluating the robustness of CSSE strategies

Here the performances of the five strategies presented in the previous Section are considered under Scenarios 2-4 of Wang et al.(2006). In Scenario 2 the phase III effect size is quite a bit (20%) lower than in phase II, and the launch threshold is set equal to the phase III effect size. In Scenario 3 the ratio between phase II and phase III effects size is the same as in the latter case, and the launch threshold is set at a value somewhat lower than δ_3. In Scenario 4 the phase III effect size is much lower (50%) than in phase II, and the launch threshold is set as in Scenario 3.

In order to evaluate the performances of the strategies their OP, which is an important evaluation tool, is computed. Then, focusing on the behavior of sample size estimators, their bias (i.e. $|M_I - E[M_n^\bullet | d_{2,n}^\bullet > \delta_{0L}]|$) and their variability must be evaluated. The MSE of sample size estimators is in particular taken into consideration - this concerns both bias and variability, since it is the sum of the square of the bias with the variance. We also report the mean of sample size estimators.

4.1 Design of the study

In a recent work we stated (De Martini, 2010) that phase II sample sizes n between $M_I/3$ and $4M_I/3$ were the most important from a practical point of view, and that higher values of n were evaluated to look at the asymptotic behavior of SSE strategies only. Now, it should be noted that under Scenarios 2-4 sample size estimation is not consistent, since $M_n^\bullet$ does not tend to M_I (but to $M_I k^2$, with $k < 1$). Moreover, $OP_n(\bullet)$ does not tend to $1 - \beta$ (but to $\pi_{\delta_3}(M_I k^2) < \pi_{\delta_3}(M_I) \simeq 1 - \beta$). Then, the behavior of SSE strategies is not relevant for high values of n either from a practical standpoint or from a theoretical one. As regards the level of the power to be achieved, Wang et al.(2006) argued that $1-\beta = 80\%$ is a low power for CSSE, and this point is confirmed in De Martini (2010). Here, under Scenarios 2-4 where $\pi_{\delta_3}(M_I k^2) < 1 - \beta$, a power choice of 80% appears even more penalizing.

Consequently, for evaluating robustness only three n settings are considered in all Scenarios (i.e. $2M_I/3$, M_I and $4M_I/3$, avoiding $n = M_I/3$ because of the poor performances of all strategies under Scenario 1), and the power is set at $1 - \beta = 90\%$.

As regards the effect size parameters, in Scenario 2 we set $\delta_3 = 0.2, 0.5, 0.8$, $k = 0.8$ (that implies $\delta_2 = \delta_3/0.8$) and $\delta_{0L} = \delta_3$; we, therefore, consider 9 settings (i.e. 3 ns × 3 δs). In Scenario 3 we set the launch threshold δ_{0L} at 0.1, with $\delta_3 = 0.2, 0.5, 0.8$; then we set $\delta_{0L} = 0.25$, with $\delta_3 = 0.5, 0.8$; as in Scenario 2, we have $\delta_2 = \delta_3/0.8$; we, therefore, consider 15 settings (i.e. 9 + 6). In Scenario 4 we set $\delta_3 = 0.2, 0.5, 0.8$, $k = 0.5$ (i.e. $\delta_2 = \delta_3/0.5$) and $\delta_{0L} = 0.1$, so that 9 settings are considered.

4.2 Results

Under Scenarios 2 and 4 all strategies do not present robust behavior. Being Scenario 3 the closest to Scenario 1, the strategies show better performances.

In scenario 2 the best performer is BAT, which shows the maximum OP (62.7%) with $\delta_3 = 0.8$ and $n = 4M_I/3$; its minimum was 55.9%, with $\delta_3 = 0.2$ and $n = 2M_I/3$. The OPs are so low that a discussion on the MSE and the average of sample size estimators is unnecessary. These poor performances are due to too high launch threshold settings with respect to true effect sizes, inducing low launch probabilities, with subsequent low OPs.

In Scenario 4 the best performer is 1SES, whose OP is around 50% with every δ_3 and n. *A fortiori* we do not discuss MSEs. Poor performances under this Scenario are caused by too large differences between the effect sizes of the two phases, i.e. $\delta_3 << \delta_2$, so that pilot data estimate a sample size (i.e. $M_I k^2$) too much lower than the interesting one (i.e. $M_I k^2 << M_I$ with $k = 0.5$). Consequently, all strategies provide small sample size estimates, inducing low OPs.

Scenario 3 is the closest to Scenario 1, where COS was clearly the best performer (De Martini, 2010). We then expect the strategies performances not to be far from those under Scenario 1. Table 1 reports OPs, together with the MSE of SS estimators and their averages in the 9 settings with $\delta_{0L} = 0.1$. All strategies are quite robust, performing better than under Scenarios 2 and 4: their OPs are higher than 70%. Nevertheless, the OPs are a bit lower than those of Scenario 1, as are the MSEs, as a consequence of the differences between phase III and phase II ES.

As it concerns strategy comparison, the (expected) better performances of COS are not as clear as under Scenario 1. It can be easily noted that COS performs better than PWS (higher OP, lower MSE, average sample size closer to M_I). As regards BAT, although the average over all 15 settings of its OPs is 3.3% larger than that of COS (which seems in any case acceptable resulting 77.5%), we still prefer COS because its average sample size is closer to M_I and it mainly presents a dramatically lower MSE: the absolute error (i.e. MSE$^{1/2}$) of BAT is more than 4 times higher than that of COS, on average over all 15 settings (i.e. the average of the rates of MSE$^{1/2}$ of BAT with respect to those of COS among the 15 settings is 4.16, in other words 316% higher). Among 3QS and 1SES, which perform similarly, we prefer the former: although the OP of 1SES is 2.0% larger, its MSE$^{1/2}$ is higher too (40%) with respect to that of 3QS, and the average sample size of 3QS is 2.2 times closer to M_I. Now the point is which is the better performer between COS and 3QS. As regards OPs, 3QS provides, on average, 81.7%, i.e. 4.2% more than COS. COS sample size estimator however performs somewhat better: the average sample size of COS is 2.0 times closer to M_I, and, mainly, the MSE$^{1/2}$ of 3QS is, on average, 2.2 times higher than COS. For these reasons we maintain that under Scenario 3 too COS is the best strategy.

5 Bias correction

In the previous Section we evaluated the robustness of CSSE strategies when a structural bias is present, resulting from differences between phase II and phase III effect sizes (see also Remark 1). This bias, as well as those present in other branches of statistical theory, can nonetheless be corrected or at least reduced. In this Section an intuitive technique of bias correction is presented, and the corrected CSSE strategies related to those recalled in Section 3 are derived.

In the context of CSSE the authors usually suggest to modify the type I and/or type II errors to reduce the bias. Fay et al. (2007), although no structural bias was considered (i.e. $\delta_2 = \delta_3$), observed that the Bayesian CSSE strategy provided an OP higher than requested, and so applied a correction to the nominal power (e.g. a power of 76% was suggested for $1 - \beta = 80\%$).

Wang et al.(2006), in presence of a structural bias, suggested techniques based on the modification of the type I and/or type II errors. One of these techniques consists in applying $\beta^* < \beta$. It is worth noting that the launch probability is not modified, whereas the correction influences sample size estimators merely. Wang et al.(2006) argue that these strategies are quite suitable under Scenarios 2-4, when applied to 1SES. For example when $\delta_2 = 0.3$ and $\delta_3 = 0.2$, if $\alpha = 0.025$ and $1 - \beta = 80\%$, then $1 - \beta^* = 92\%$ is suggested. Note that their conclusions are based on the computation of the average of sample size estimators only (see Table 4 in their work).

5.1 Postulating the correction for the effect size

Our proposal concerns a correction to be applied directly on the effect size. In practice, the corection should be applied to the effect size estimator $d^{\bullet}_{2,n}$. This idea stems from two observations: on one hand, CSSE aims at studying the performances of $M^{\bullet}_n$, based on $d^{\bullet}_{2,n}$, which is, under Scenarios 2-4, a biased (and non consistent) estimator of δ_3; on the other hand, the structural bias can be modeled through $\delta_2 = \delta_3/k$. Hence, to improve CSSE performances we find it natural and intuitive to speculate first about k. Being k_c the postulated correction, $M^{\bullet}_n$ can be directly modified by using $d^{\bullet}_{2,n} \times k_c$, instead of $d^{\bullet}_{2,n}$ only, in its formula. We then obtain ${}_cM^{\bullet}_{2,n}$. Consequently, the assumption on $M_{\max}$ changes, and the new maximum is ${}_cM_{\max} = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/(\delta_{0L} * k_c)^2 \rfloor + 1$.

Finally, note that the launch probability is not modified by the correction suggested here, and this is in accordance with those proposed by Wang et al.(2006). Indeed, $d^{\bullet}_{2,n}$ remains the same, $M^{\bullet}_n$ alone is modified.

5.2 Corrected CSSE strategies

Corrected fixed-γ strategies (viz. PWS, 1SES, 3QS). The γ conservative estimator of δ_3 is multiplied by the correction k_c, obtaining ${}_cd^{\gamma}_{2,n} = (d_{2,n} - z_\gamma/\sqrt{n/2}) * k_c$, and, consequently, ${}_cM^{\gamma}_n = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/({}_cd^{\gamma}_{2,n})^2 \rfloor + 1$.

Corrected COS. We first refer to the corrected version of the estimated OP, which, following from (3), is:

$$ {}_c\hat{O}P_n(\gamma) = \sum_{m=2}^{{}_cM_{\max}} \{\Phi(\sqrt{\frac{n}{2}}(\frac{z_{1-\alpha}+z_{1-\beta}}{k_c\sqrt{(m-1)/2}} - d_{2,n}) + z_\gamma) - \Phi(\sqrt{\frac{n}{2}}(\frac{z_{1-\alpha}+z_{1-\beta}}{k_c\sqrt{m/2}} - d_{2,n}) + z_\gamma)\}\Phi(\sqrt{\frac{m}{2}}k_c d_{2,n} - z_{1-\alpha}) \tag{4} $$

Then, being $g_{O,n}$ the argument of the constrained maximum of (4), we make use of ${}_cd^{g_{O,n}}_{2,n} = (d_{2,n} - z_{g_{O,n}}/\sqrt{n/2}) * k_c$ to compute the sample size estimate ${}_cM^{g_{O,n}}_n$.

Corrected BAS. k_c modifies Bayesian strategies through the corrected power ${}_c\pi_{Ba_n}(m) = \int_{-\infty}^{\infty} \pi_{z*k_c}(m)\phi_{d_{2,n},2/n}(z)\,dz$, so that ${}_cM^{BaS}_n = \min\{m \,|\, {}_c\pi_{Ba_n}(m) > 1-\beta\}$, and ${}_cM^{BaT}_n = \min\{{}_cM^{BaS}_n, {}_cM_{\max}\}$.

Remark 4. From the mathematical perspective, the correction techniques of Wang et al.(2006) and that we proposed here are connected. In fact, both approaches aim to correct $\lfloor 2(z_{1-\alpha}+z_{1-\beta})^2/(\delta_{2,n}^{\gamma})^2\rfloor+1$ in order to make it closer to M_I. Corrections can be brought to the numerator (Wang et al., 2006) or to the denominator (as we suggest). For each correction on the numerator there exists an equivalent correction for the denominator: for example, for every $\beta^* < \beta$ there exists $k_c < 1$ such that $z_{1-\alpha}+z_{1-\beta^*} = (z_{1-\alpha}+z_{1-\beta})/k_c$. For this reason we will not include correction techniques based on the modification of the type I and/or type II errors in the ensuing comparison (Section 6).

Remark 5. Note that since $k_c < 1$ the average and the standard deviation of ${}_cM_n^{\gamma}$ are increased by $1/k_c$ times with respect to those of M_n^{γ}. As a consequence, $MSE[{}_cM_n^{\gamma}]$ is often higher than $MSE[M_n^{\gamma}]$. We expect analogous behavior of the corrected COS and BAT.

Remark 6. If the postulated correction k_c is right (i.e. $\delta_2 = \delta_3/k_c$), then all the considered strategies are consistent (i.e. ${}_cM_{2,n}^{\bullet}$ tends to M_I). We do not nevertheless fall under Scenario 1: although the mean of ${}_cd_{2,n}$ is actually δ_3, its variance (i.e. $2k_c^2/n$) is different from that of $d_{2,n}$ under Scenario 1 (i.e. $2/n$).

6 A study comparing corrected CSSE strategies

We here evaluate the performances of the five corrected strategies introduced in Section 5.

6.1 Design of the study

Scenario 3 alone of the previous Section 4 is examined, with the 9 settings where $\delta_{0L} = 0.1$. We recall that $k = 0.8$. As regards k_c, there are three postulated corrections: a smaller than necessary one, i.e. $k_c = 0.9$, the right one, i.e. $k_c = 0.8$, and a higher than necessary one, i.e. $k_c = 0.7$. So, 27 settings are evaluated (3 ns × 3 δ_3s × 3 k_cs). Since our CSSE strategies are consistent when $k_c = 0.8$, we also evaluate their asymptotic behavior, with n up to $4M_I$, for the three δ_3s. Once again, we look at OPs of CSSE strategies, and at the MSEs and averages of their sample size estimators.

6.2 Results

Table 2 reports OP, MSE and average of sample size estimators obtained with $\delta_3 = 0.5$ alone, i.e. under 9 settings out of 27.

When $\delta_3 < \delta_2$, applying a correction $k_c < 1$ improves the OPs of all strategies, but also increases their MSEs. The averages of sample size estimators increase too, and the bias augments in all circumstances but one (i.e. for PWS with $k_c = 0.9$).

With $k_c = 0.9$, the OPs improve, on average over the 9 settings, from 4% (1SES) to 6.8% (PWS), but they are still lower than 90% (from 80.1% of PWS to 88.5% of 1SES). The MSE$^{1/2}$ of corrected strategies increases, on average, with respect to the uncorrected ones, from 16% with PWS (i.e. MSE$^{1/2}$ of corrected PWS is 1.16 times higher than that of PWS without any correction) to 30% of 1SES. Hence, the correction, although smaller than necessary, works well because the little increase in MSEs is counter balanced by the good improvements in OPs for all strategies. In particular, the OP of COS increases 6.1%, on average, becoming 83.7% (3QS, 87.1%). In practice, the corrected COS provides the second best improvement in OP and the second smallest increase in MSE$^{1/2}$ (19%). Moreover, the MSE$^{1/2}$ of COS is still the lowest: those of PWS, 3QS, 1SES and BAT are, on average, 1.28, 2.53, 3.53 and 5.04 times higher, respectively. Since COS was the best performer under Scenario 3, it remains the recommended strategy even with a smaller than necessary correction.

With a higher than necessary correction, i.e. $k_c = 0.7$, the OPs of all strategies result higher than 90%, going, on average, from 92.3% (PWS) to 95.1% (BAT) (COS, 94.0%). Nevertheless, the MSE$^{1/2}$ increases, on average, from 2.14 times with PWS to 2.46 times with COS, with respect to those obtained without correction. Also, the averages of sample size estimators are much higher than M_I. These performances are caused by too high sample size estimates provided by this excessive correction. In the light of sample size estimator behaviors, it would be preferable to avoid higher than necessary corrections. As regards detailed results, the MSE$^{1/2}$ of COS is once again the lowest: the enlargements of the other strategies go from 12% of PWS to 331% of BAT.

With the right correction of $k_c = 0.8$, all strategies converge, as n tends to ∞: their OPs tend to 90% and the MSEs tend to zero. (see Figures 1 and 2). It should also be noted that when $2M_I \leq n \leq 4M_I/3$, all strategies provide OPs closer to 90% than those obtained with $k_c = 0.9$. Furthermore, considering these pilot sample sizes of practical interest, COS converges faster. Indeed, the strategy with averaged OP (over the usual 9 settings) closest to 90% is COS (89.4%), where the highest differences are provided by 1SES (91.8%). The MSE$^{1/2}$ increases, as expected, from 51% of PWS to 74% of 1SES, with respect to that provided by the strategies without correction. Also, the MSE of COS is the lowest: the enlargements in MSE$^{1/2}$ go from 21% of PWS to 4.75 times (i.e. 375%) of BAT. Once again COS is recommended, and this was expected, since postulating the right correction makes estimations quite close to those under Scenario 1.

Finally, we emphasize that the OPs provided by all corrected strategies are a little higher (and closer to 90%) than the respective ones computed in De Martini (2010) under Scenario 1, i.e. where correction was not needed (see Figure 1). Moreover the MSEs with $k_c = 0.8$ are comparable to those under Scenario 1 (see Figure 2). These behaviors are due to higher launch probabilities with respect to Scenario 1 (due to $\delta_2 > \delta_3$), and to quite close sample size estimation.

7 Discussion

The statistical literature contains many works on SSE for planning a generical experiment, as well as, in particular, a randomized controlled trial. Over the last five years, the conservative approach to SSE, which takes into account the variability of phase II pilot data, has been proposed under some different approaches. Chuang-Stein (2006) and Fay et al. (2007) argued a Bayesian approach, which consists in averaging the estimated power on the basis of the posterior distribution of the effect size. In the frequentist framework, Wang et al. (2006) suggested a simple strategy consisting in the estimation of the sample size on the basis of the 1 standard error conservative estimate of the effect size. However, all these authors did not put sufficient emphasis on the variability of sample size estimators, for example by computing their MSEs, and we believe that this point is mandatory.

Recently, we compared (De Martini, 2010) some Bayesian and some frequentist CSSE strategies with a new calibrated optimal γ conservative one (viz. COS). The Overall Power (OP) of phases II and III, together with the average of sample size estimators and, mainly, at their MSE, were considered. The indispensable launch threshold criterion, introduced by Wang et al.(2006), was adopted, and it was assumed that the effect size in phase II was the same as that of phase III (i.e. the so-called Scenario 1). COS resulted the best strategy, with good OPs, and with MSEs considerably lower than those of the other strategies. Nevertheless, the size of the phase II sample used for CSSE should be around the ideal phase III sample size, the launch threshold should be lower than one half of the true phase III effect size, and the prefixed power should be set at 90%. These three assumptions were considered as operating CSSE conditions.

In practice, small deviations from Scenario 1 would not substantially change these results. Nevertheless, in some circumstances the difference between phase II and phase III effect sizes can be not negligible, for example in the presence of a more restrictive inclusion criteria in phase II, or even when a lack of knowledge of dose-response relationship occurs. In these cases, phase III effect size is often lower than phase II. Consequently, not only CSSE becomes inconsistent for large samples, but it might also provide bad results for finite phase II sample sizes of practical interest. Wang et al.(2006) considered some of these situations (viz. Scenarios 2-4), where CSSE was applied in the presence of bias.

In this work, we have compared the performances of some Bayesian and some frequentist CSSE strategies with COS under Scenarios 2-4, in order to evaluate whether COS would still have provided good results. The results were mainly based on OP and MSE. As regards Scenarios 2 and 4, COS, unlike under Scenario 1, did not improve the performances of fixed-γ or of Bayesian estimators, and we confirmed the results in Wang et al. (2006), where all strategies performed poorly. In Scenario 2 there was the problem of a too high launch threshold, so we confirm that it would be better to set $\delta_{0L} < \delta_3/2$. In Scenario 4 poor performances were due to the high difference between the effect sizes of phase II and phase III. COS was still the best performer (thanks to its low MSE) under Scenario 3, although its supremacy is not as clear as in Scenario 1. Assuming the continuity of the performances of different strategies when the ratio k between δ_2 and δ_3 changes, in the light of results under Scenarios 1 and 3 we find that for small differences between δ_2 and δ_3 (i.e. $\delta_3/\delta_2 \geq 0.8$) COS remains the best strategy.

In practice, COS is a very good strategy when CSSE is consistent, that is under Scenario 1 when $\delta_2 = \delta_3$: in this case COS converges quickly, and with small variability. Consequently COS suffers when a structural bias is present (i.e. $\delta_2 \neq \delta_3$), and consequently a wrong sample size is being estimated. Nevertheless, if differences between effect sizes are small, i.e. lower than 20%, COS still performs well, provided that pilot sample sizes are not large (i.e. $n \leq 4M_I/3$) in order to avoid the wrong asymptotic convergence.

Now, it should be noted that there exist some techniques for correcting bias in CSSE. Wang et al. (2006) suggested correcting α or/and β levels, in accordance with some other works in the field. Here, we proposed to apply the correction directly on the source of the bias, that is on the estimation of δ_3. Indeed, in our opinion, to correct α and/or β is more complicated and counter-intuitive than directly correcting $d^{\bullet}_{2,n}$, i.e the estimator of δ_3. In particular, when the error levels are modified a check of the amplitude of the correction is missing. On the contrary, $d^{\bullet}_{2,n}$ can simply be modified by expressing the postulated correction k_c. Hence, if k_c is right, then CSSE is consistent and the structural bias is, asymptotically, corrected; otherwise, if k_c is at least close to k, then this bias can be reduced, specifically for finite pilot sample sizes of practical interest. In both situations, the framework of CSSE returns to being close to Scenario 1, so that the peculiarity of COS can be exploited.

We, then, evaluated the behavior of k_c-corrected CSSE under Scenario 3, i.e. when the phase III effect size is 20% lower than phase II. When the postulated correction is the right one, COS clearly performs better than the other strategies, providing good OPs and the smallest MSEs. It is worth noting that in this case the OPs of COS are even higher than those observed under Scenario 1 in De Martini (2010): this is due to a higher launch probability, since $\delta_2 > \delta_3$, and to a similar estimation performance. Hence, paradoxically, the bias can be exploited for improving CSSE, should the right correction be applied. When the postulated correction is smaller than necessary COS is still the best performer, and it works quite well in general. A higher than necessary correction should be avoided.

8 Conclusions

We maintain, in the presence of structural bias too, the operating conditions for CSSE to be: the launch threshold should be set lower than one half the true phase III effect size; the pilot sample should be around the ideal one (i.e. $n \geq 2M_I/3$); a power of 90% should be adopted.

As a consequence of phase III effect sizes lower than phase II, low OPs often occur. If the difference between δ_2 and δ_3 is quite small (i.e. $1 \geq \delta_3/\delta_2 \geq 0.8$), then CSSE can still provide acceptable results, mainly by adopting COS. Nevertheless, a correction to CSSE can be applied, and it affects only sample size estimates, not launch probabilities. If a good correction is applied, i.e. close to the right one, and preferably a bit smaller rather than a bit larger, all CSSE strategies provide improved results, and COS results the best. If the right correction is applied, COS works even better than in the absence of structural bias.

References

Chuang-Stein, C. (2006). Sample Size and the Probability of a Successful Trial. *Pharmaceutical Statistics* 5: 305–309.

De Martini, D. (2010). Adapting by calibration the sample size of a phase III trial on the basis of phase II data. *Pharmaceutical Statistics*. Published online 10 Feb.

Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fay, M.P., Halloran, M.E., Follmann, D.A. (2007). Accounting for Variability in Sample Size Estimation with Applications to Nonhaderence and Estimation of Variance and Effect Size. *Biometrics* 63: 465–474.

Hung, H.M.J., Wang, S.J., O'Neill, R.T. (2006). Methodological issues with adaptation of clinical trial design. *Pharmaceutical Statistics* 5: 99–107.

Rosner, B. (2005). *Fundamentals of Biostatistics*. 6th ed. Boston: Duxbury Press.

Shao, J., Chow, S.C. (2002). Reproducibility probability in clinical trials. *Statistics in Medicine* 21: 1727–1742.

Wang, S.J., Hung, H.M.J., O'Neill, R.T. (2006). Adapting the sample size planning of a phase III trial based on phase II data. *Pharmaceutical Statistics* 5: 85–97.

Table 1	OP, MSE and average of sample size estimators					
Overall Power	n	PWS	3QS	1SES	COS	BAT
$\delta_3 = 0.2$	$2M_I/3$	70.67%	75.04%	72.85%	75.31%	80.19%
	M_I	72.68%	79.95%	80.52%	77.42%	80.17%
	$4M_I/3$	73.37%	81.63%	83.79%	77.55%	79.53%
$\delta_3 = 0.5$	$2M_I/3$	73.08%	83.03%	85.48%	79.49%	82.50%
	M_I	73.69%	83.75%	87.47%	78.13%	81.09%
	$4M_I/3$	73.88%	83.44%	87.35%	77.20%	79.97%
$\delta_3 = 0.8$	$2M_I/3$	73.75%	84.05%	87.12%	78.96%	82.95%
	M_I	74.26%	84.30%	88.22%	77.68%	81.49%
	$4M_I/3$	74.45%	83.80%	87.73%	76.92%	80.40%
MSE of SS Est.	n	PWS	3QS	1SES	COS	BAT
$\delta_3 = 0.2$	$2M_I/3$	105405	173763	232528	34649	423681
	M_I	78751	130186	181341	35620	227369
	$4M_I/3$	62875	95904	137493	38027	131601
$\delta_3 = 0.5$	$2M_I/3$	12503	45324	78144	3971	229328
	M_I	3647	16275	32738	3349	58581
	$4M_I/3$	1815	7154	15497	2543	14208
$\delta_3 = 0.8$	$2M_I/3$	3207	17894	34876	1055	164605
	M_I	544	4766	13690	719	31270
	$4M_I/3$	272	1060	3849	474	4631
Av. of SS Est.	n	PWS	3QS	1SES	COS	BAT
$\delta_3 = 0.2$	$2M_I/3$	428.7	615.4	715.8	407.0	742.2
	$M_I = 526$	408.3	578.9	675.9	423.8	600.3
	$4M_I/3$	392.3	543.2	633.6	423.7	519.4
$\delta_3 = 0.5$	$2M_I/3$	81.3	145.7	192.9	92.2	233.3
	$M_I = 85$	69.0	111.4	144.4	83.9	127.2
	$4M_I/3$	63.9	94.8	118.3	75.7	91.0
$\delta_3 = 0.8$	$2M_I/3$	33.1	65.2	92.0	38.9	137.7
	$M_I = 33$	27.2	45.7	62.5	33.7	57.8
	$4M_I/3$	25.2	37.2	47.5	29.6	36.7

Table 1. Performances of different strategies under Scenario 3, with $\delta_{0L} = 0.1$.

Table 2	OP, MSE and average of sample size estimators					
Overall Power	n	PWS	3QS	1SES	COS	BAT
$k_c = 0.9$	$2M_I/3$	79.49%	87.54%	88.97%	84.66%	87.17%
	M_I	80.51%	88.78%	91.50%	84.09%	86.54%
	$4M_I/3$	80.97%	88.85%	91.80%	83.71%	85.94%
$k_c = 0.8$	$2M_I/3$	85.81%	91.54%	91.90%	89.53%	91.48%
	M_I	87.09%	93.13%	94.79%	89.68%	91.48%
	$4M_I/3$	87.71%	93.48%	95.40%	89.73%	91.32%
$k_c = 0.7$	$2M_I/3$	91.48%	94.71%	94.08%	93.81%	95.10%
	M_I	92.79%	96.44%	97.13%	94.39%	95.48%
	$4M_I/3$	93.42%	96.91%	97.89%	94.65%	95.58%
MSE of SS Est.	n	PWS	3QS	1SES	COS	BAT
$k_c = 0.9$	$2M_I/3$	19280	72276	124432	6780	356741
	M_I	5216	26411	53015	5451	91625
	$4M_I/3$	2189	11746	25536	3847	22324
$k_c = 0.8$	$2M_I/3$	32249	121648	208530	13047	583438
	M_I	8861	45782	90724	10315	151353
	$4M_I/3$	3658	21134	44867	7164	37806
$k_c = 0.7$	$2M_I/3$	58518	217829	371002	26981	1014780
	M_I	17304	84775	164929	21440	266519
	$4M_I/3$	7889	40968	83939	15181	69011
Av. of SS Est.	n	PWS	3QS	1SES	COS	BAT
$k_c = 0.9$	$2M_I/3$	100.3	179.8	238.0	114.0	287.0
	$M_I = 85$	85.0	137.4	178.2	103.6	156.9
	$4M_I/3$	78.8	116.9	145.9	93.5	112.3
$k_c = 0.8$	$2M_I/3$	126.8	227.4	301.1	144.4	364.3
	$M_I = 85$	107.5	173.8	225.4	131.2	198.5
	$4M_I/3$	99.5	147.9	184.6	118.4	142.0
$k_c = 0.7$	$2M_I/3$	165.5	296.8	393.1	188.7	475.7
	$M_I = 85$	140.3	226.8	294.2	171.5	259.1
	$4M_I/3$	129.9	193.0	240.9	154.7	185.3

Table 2. Performances of corrected strategies under Scenario 3, with $\delta_{0L} = 0.1$, and with $\delta_3 = 0.5$.

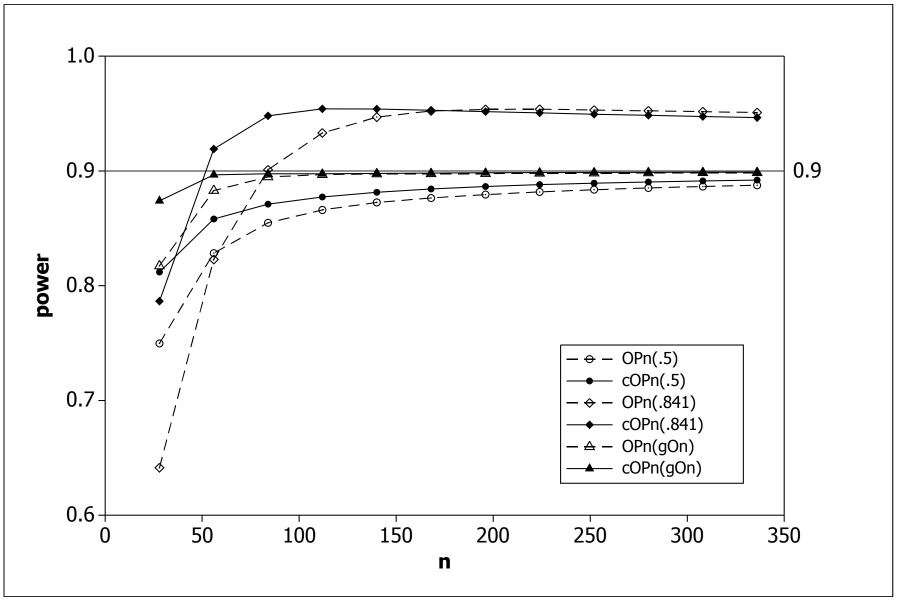


Figure 1. Overall Powers of PWS, 1SES and COS under Scenario 3 with the right correction $k_c = 0.8$ and under Scenario 1, with $\alpha = 0.025$, $1 - \beta = 0.9$, $\delta = 0.5$ and $\delta_{0L} = 0.1$.

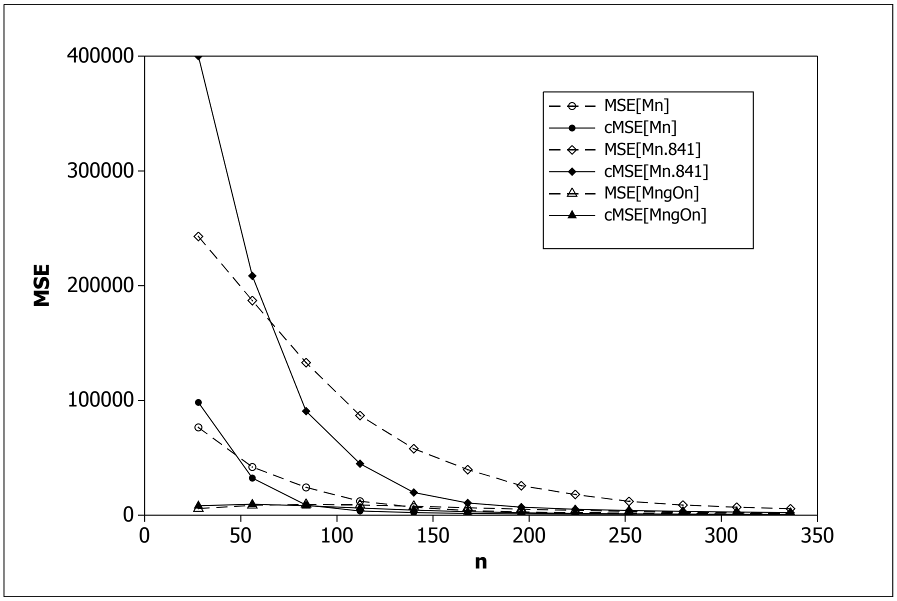
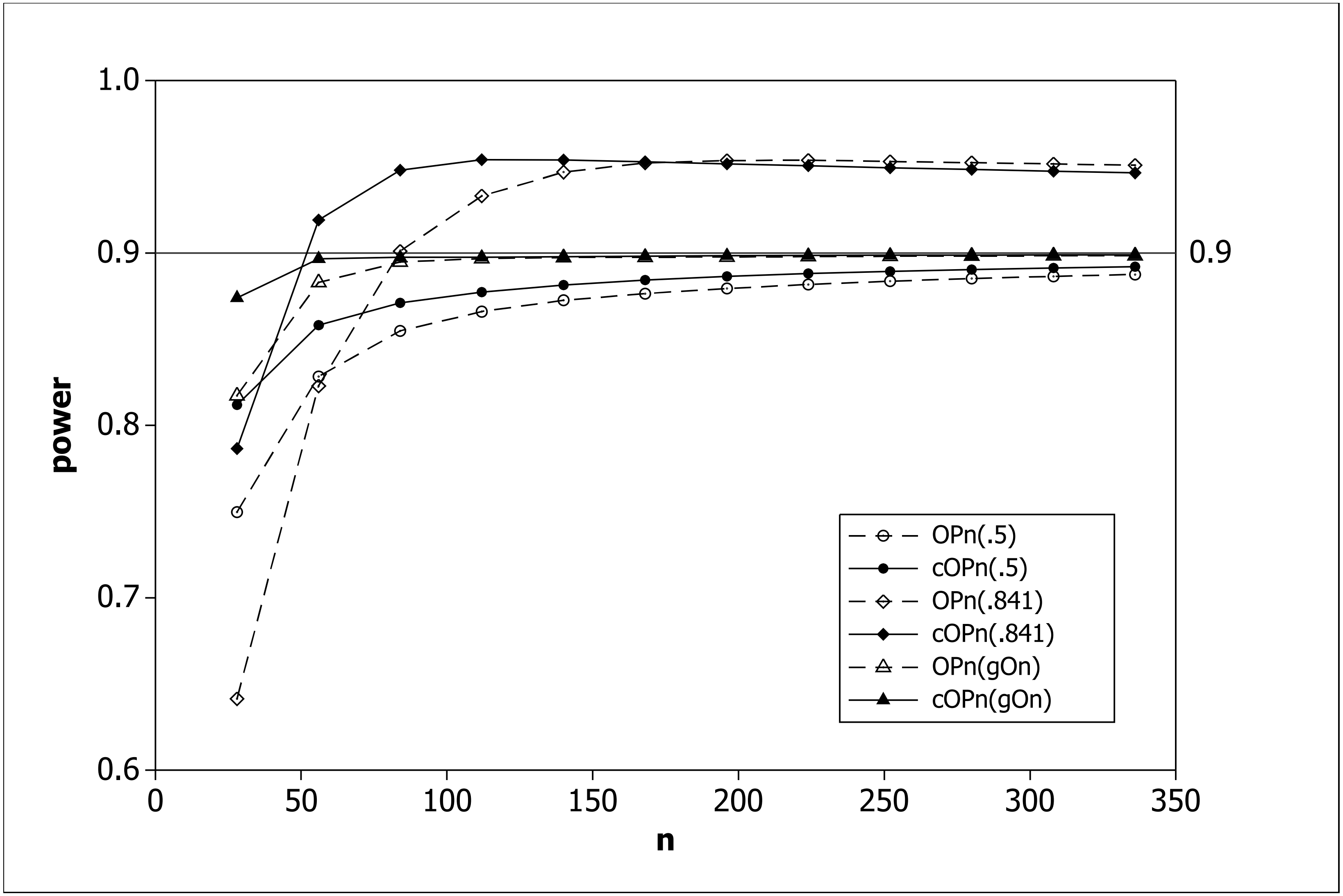
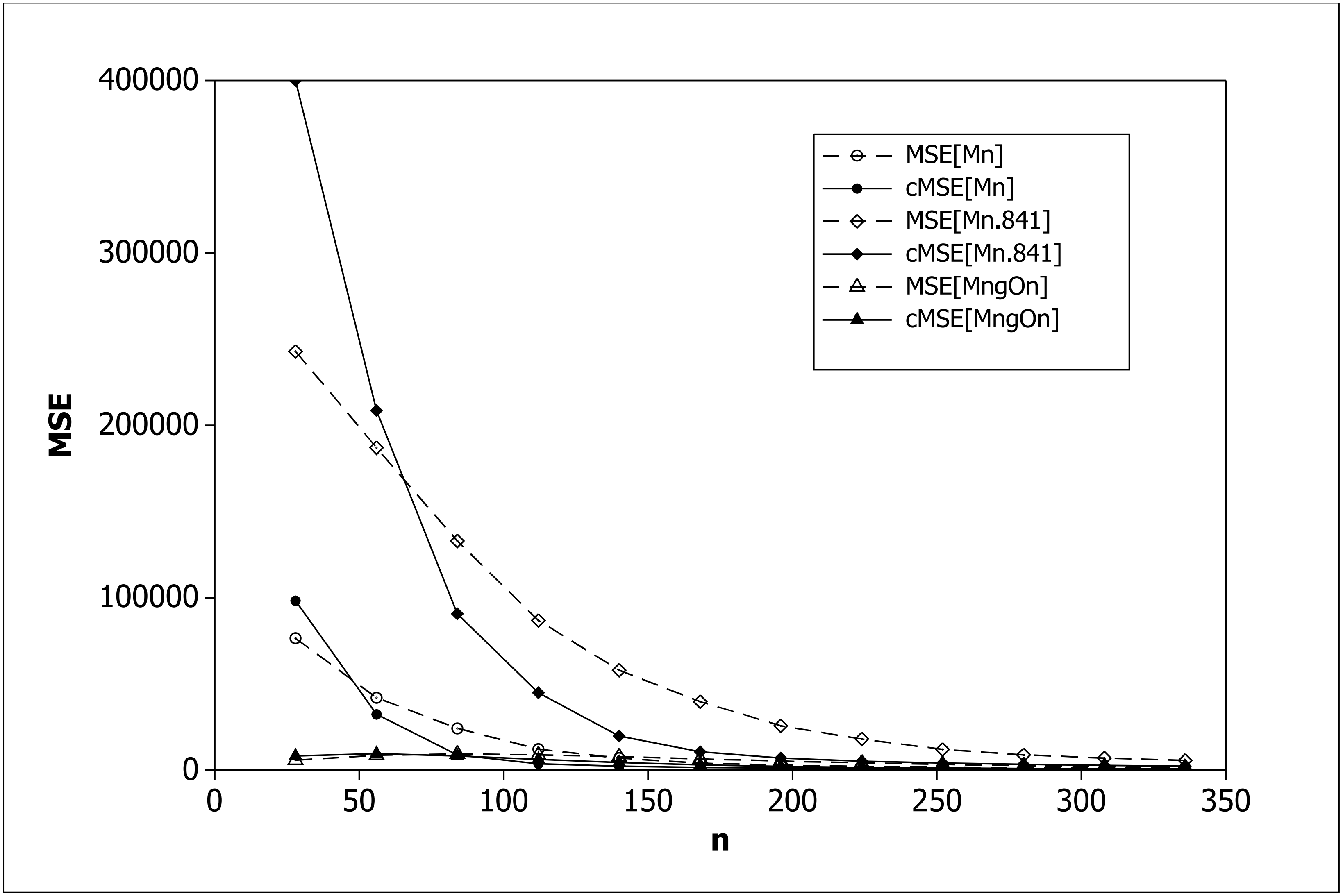


Figure 2. MSEs of the sample size estimators of PWS, 1SES and COS under Scenario 3 with the right correction $k_c = 0.8$ and under Scenario 1, with $\alpha = 0.025$, $1 - \beta = 0.9$, $\delta = 0.5$ and $\delta_{0L} = 0.1$.


1.0
0.9
0.8
0.7
0.6
power
0
50
100
150
200
250
300
350
n
0.9
OPn(.5)
cOPn(.5)
OPn(.841)
cOPn(.841)
OPn(gOn)
cOPn(gOn)


400000
300000
200000
100000
0
MSE
0
50
100
150
200
250
300
350
n
MSE[Mn]
cMSE[Mn]
MSE[Mn.841]
cMSE[Mn.841]
MSE[MngOn]
cMSE[MngOn]