Rapporto n° 242

Allocating the Sample Size in Phase II and III Trials to Optimize Success Probability

Daniele De Martini

Ottobre 2013

Dipartimento di Statistica e Metodi Quantitativi
Università degli Studi di Milano Bicocca
Via Bicocca degli Arcimboldi 8 - 20126 Milano - Italia
Tel +39/02/64483102/3 - Fax +39/2/64483105
Segreteria di redazione: Andrea Bertolini

Allocating the Sample Size in Phase II and III Trials to Optimize Success Probability

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

Summary Clinical trials of phase II and III often fail due to poor experimental planning. Here, the problem of allocating available resources, in terms of sample size, to phase II and phase III is studied with the aim of increasing success probability. The overall success probability (OSP) is accounted for, where it is assumed that phase II data are not considered for confirmatory purposes and that are used for planning phase III through sample size estimation. Being r the amount of resources allocated to phase II, $OSP(r)$ is a concave function and there exists an optimal allocation r_{opt}. If M_I is the sample size giving the desired power to phase III, and kM_I is the whole sample size that can be allocated, it is then indicated how large k and r should be in order to achieve levels of OSP of practical interest. For example, when 5 doses are evaluated in phase II and 2 parallel phase III confirmatory trials (one-tail type I error $= 2.5\%$, power $= 90\%$) are considered with 2 groups each, $k = 24$ is needed to obtain $OSP \simeq 75\%$, with $r_{opt} \simeq 50\%$. When k is large enough, r_{opt} is close to 50%. The choice of k depends mainly on how many phase II treatment groups are considered, not on the effect size of the selected dose. To increase OSP, phase III sample size may be estimated conservatively: an improvement of $\simeq 3\%$ is given when $OSP \simeq 75\%$ and optimal conservativeness is adopted. Resources larger than those usually employed should be allocated to phase II to increase OSP.

Keywords: Launching rules; sample size estimation; overall success probability; optimal allocation; sufficient resources.

1 Introduction

It is common knowledge that phase II clinical trials aim is mainly exploratory, while that of phase III is confirmatory, and that phase II also serves to enhance planning for the subsequent phase III. Usually, phase II is small with respect to phase III - on average, the phase II sample size is about 25% of that of phase III. The rate of trials failure, which is around 60% and 40% for phase II and III, respectively, suggest that this habit might be not helpful. In general, low success probabilities are often due to low sample size (see [1], also reporting the above mentioned data). Here, we study sample size problems from the perspective of a drug development project, which means considering jointly phase II and phase III sample sizes.

To introduce the problem, by way of example, let us suppose that a phase II trial with 2 parallel arms has been run and that a phase III with the same design needs to be planned - phase II often involves more than 2 groups and phase III consists of 2 simultaneous trials. Assume that the minimum efficacy value (i.e. standardized effect size) that should be observed to then launch phase III is 0.15 and that a slightly higher value than this has been observed in phase II, so that phase III has now to be launched. With one-sided $\alpha = 2.5\%$ and power $1 - \beta = 90\%$, approximately 940 patients should be recruited for each of the two groups of a parallel design, if the observed effect size is adopted for sample size computation - this is the so-called pointwise estimation strategy, which is one of the available sample size estimation strategies [2, 3]. This number (940) is quite high, but not beyond the range of those usually adopted in phase III trials (visit, for example, *clinicaltrials.gov*, a service of the U.S. NIH). So, assume that the research team decided to actually launch phase III. Moreover, assume that 60 patients per group have been recruited in phase II, so that the total number of patients enrolled in phase II and phase III is about 2000.

Now, the point is: if the resources for studying these 2000 patients were actually available, would there be an allocation of sample size better than 60/940? Would, for example, 400 data allocated to phase II and, at most, 600 to phase III has been a better choice? It is worth noting that we wrote *at most* because once 400 data per group come from phase II, the phase III sample size computed on the basis of the latter data is not necessarily 600, where it is almost surely lower. Moreover, what does "better allocation" mean? And, is there an optimal allocation?

Besides dose selection and safety evaluation, phase II aims are that of correctly deciding go/no-go, that of launching phase III (i.e. go) with a high probability if a meaningful effect really exists, and that of well estimating the drug effect size to indicate a phase III sample size (M) as close as possible to the ideal one (i.e. the one providing the desired probability of success of phase III); the aim of phase III is to prove efficacy with a high probability, once again if a meaningful effect really exists.

Hence, the aim is to succeed with high probability in both phases, whenever the drug under study actually works well. Then, in this paper we study sample size resource allocation in terms of overall probability of success (OSP). In particular, we focus on the amount of resources that should be provided to phase II and III trials so as to attain a good level of OSP, and on how these resources should be allocated between the two phases to optimize OSP. It is assumed that phase II data provide information for phase III planning and are not used for phase III confirmatory analysis.

Analogous computations on success probability have recently been proposed by Jiang [4] under the Bayesian framework. Here, the frequentist approach is adopted; this is also due to poor performances of Bayesian sample size estimators (proposed, for example, by Chuang-Stein [5]) in terms of high variability of their results [3].

1.1 Contents

The theoretical framework is stated in Section 2, where phase II and phase III tools (mainly launching rules and sample size definitions, respectively) are shown. In Section 3, OSP is presented and some formulas are given. The results on the behavior of OSP are shown in Section 4. In Section 5 the problem of the whole amount of resources needed is studied. The variability of phase III sample size, which is estimated on the basis of phase II data, is discussed in Section 6 in terms of mean and MSE. The impact of conservative sample size estimation on this allocation problem is evaluated in Section 7. The final discussion follows in Section 8.

2 Theoretical framework

2.1 Overview

It is assumed that a certain disease is under study, and that h doses of a new drug for the disease of interest are evaluated in a phase II trial (h often varies from 3 to 7). Also, a placebo arm is run. A classical parallel design is applied in the exploratory phase II, with $h{+}1$ groups. If phase II results are promising, a single dose D is chosen and 2 phase III trials comparing to placebo are run, once again under parallel design. It is also assumed that the three trials (1 phase II and 2 phase III) share the same response variable and the same patient population, meaning that the effect size of the elected dose is the same in both phases. These assumptions allow simple sample size estimation, with no need of further adjustments such as those in [6], and are similar to the assumptions in Jiang [4], where $h = 1$ and only one phase III trial were considered. Here, all trials are run under balanced sampling, i.e. the sample size of the groups under treatment is equal to that of the placebo group.

A certain, limited, amount of resources is available to develop phase II and III trials, and this translates into a total of *at most* w patients that can be studied and should be allocated in phase II and phase III. Let $r \in (0, 1)$ be the rate of w allocated to phase II: this implies that if a sample of size n is studied for each treatment in phase II, then $n(h + 1) = rw$. So, the sample size n of each group in the phase II trial is, approximately $rw/(h + 1)$. Consequently, the whole sample size available for phase III is $w(1 - r)$, which is not used entirely (almost surely). Indeed, the phase III sample size actually adopted for each group is estimated on the basis of phase II results, and is consequently a random variable, call it M_n, such that $M_n \leq w(1-r)/4$, since phase III groups are 4 (see Section 2.3).

2.2 Phase II Tools

Let μ_D and μ_P be the means of response variables of the populations under D and under placebo respectively. Also, let σ the common standard deviation of the two populations, so that the generic standardized effect size is $\delta = (\mu_D - \mu_P)/\sigma$ - without loss of generality σ is set equal to 1. The true, unknown, effect size is δ_t. Moreover, $\bar{X}_{D,n}$ and $\bar{X}_{P,n}$ are the means of measurements of samples of size n from the two populations, and $d_n = \bar{X}_{D,n} - \bar{X}_{P,n}$ is the pointwise estimator of δ_t.

Now, call $\mathcal{L}$ the random event representing the success of phase II, where $\mathcal{L}$ stands for *launch* of phase III. $\mathcal{L}$ can be defined in some different ways: on the basis of statistical significance with an appropriate phase II type I error probability α_{II} (i.e. $\mathcal{L} \Leftrightarrow T_n > z_{1-\alpha_{II}}$); on the basis of the constraint given by a maximum sample size $m_{\max}$ for phase III (i.e. $\mathcal{L} \Leftrightarrow M_n \leq m_{\max}$); on the basis of the observed effect size overcoming a threshold of clinical relevance δ_{0L} (i.e. $\mathcal{L} \Leftrightarrow d_n > \delta_{0L}$). Kirby et al.[7] evaluated some further launching rules, which, through simple algebra, can be reduced to $\mathcal{L} \Leftrightarrow d_n > \delta_{0L}$, for some values of δ_{0L}. In [1] (Ch.3), it is shown that the three launching criteria described above can be set in order to result mathematically equivalent.

Although the launching rule based on δ_{0L} is the most intuitive, and one of the most used, let us focus on that based on $m_{\max}$. In this framework constrained by w and modeled by r, a refinement is necessary for the maximum allowed phase III sample size. Once $m_{\max}$ (i.e. the maximum phase III sample size per group) has been fixed, an intersection is needed between the two constraints for M_n (i.e. $w(1-r)/4$ and $m_{\max}$): the actual launching rule, then, becomes $M_n \leq m_{\max}(r) = \min\{m_{\max}, w(1 - r)/4\}$. This launching criterion is considered here, which allows to ease computational formulas as in [3].

For completeness, $M_n \leq m_{\max}(r)$ translates into $\mathcal{L} \Leftrightarrow d_n > \delta_{0L}(r) = \max\{\delta_{0L}, 2(z_{1-\alpha} + z_{1-\beta})\sqrt{2/w(1 - r)}\}$ (of course, the threshold for d_n must remain of a certain clinical interest - we come back to this point in Section 4.2).

The success of phase II is considered the launching of phase III. Consequently, phase II success probability is:

$$SP_{II}(r) = P_{\delta_t}(\mathcal{L}) = P_{\delta_t}(d_n > \delta_{0L}(r)) = P_{\delta_t}(M_n \leq m_{\max}(r))$$

2.3 Phase III tools

The so called Z-test is adopted, comparing the means of two normal distributions with known variance. One-sided alternatives only are taken into account, so that the statistical hypotheses are $H_0 : \mu_D = \mu_P$ and $H_1 : \mu_D > \mu_P$.

Being $\bar{X}_{D,m}$ and $\bar{X}_{P,m}$ the means of the treatment and placebo group, respectively, with samples of generic size m, let $T_m = \sqrt{m/2}(\bar{X}_{D,m} - \bar{X}_{P,m})$ be the test statistic. Given the type I error probability α, statistical significance (i.e. *a trial success*) is found if $T_m > z_{1-\alpha}$, where $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ and Φ is the distribution function of the standard normal. The power function of a generic phase III trial is: $P_\delta(T_m > z_{1-\alpha}) = \Phi(\delta\sqrt{m/2} - z_{1-\alpha}) = \pi_\delta(m)$, and, being δ_t the true and unknown value of δ, the true power is $\pi_{\delta_t}(m)$, which was simply called success probability in [1], Ch.3, i.e. $SP(m)$.

Now consider $1-\beta$ to be the desired power to be achieved in each phase III trial (e.g. 90%), where β is the type II error probability. Then, the ideal sample size per group for each phase III trial is:

$$M_I = \min\{m \,|\, SP(m) > 1-\beta\} = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/\delta_t^2 \rfloor + 1 \tag{1}$$

Ideally, an infinite number of patients should run phase II in order to obtain that the effect size estimated through phase II data coincides with δ_t (almost sure), and consequently phase III is launched with M_I data per group. In practice, n data per group are allocated in phase II, and once phase II has succeeded, i.e. conditionally to $\mathcal{L}$, phase III is run with the sample size estimated by phase II data. Several sample size estimation strategies can be adopted (see [1]). Here, M_I is estimated by the pointwise estimator of the sample size, based on the pointwise estimator of the effect size (i.e. d_n):

$$M_n = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/d_n^2 \rfloor + 1 \tag{2}$$

The adoption of the pointwise strategy is made for simplicity and also because the performances of M_n, in terms of OSP and MSE are acceptable, although not best [3]. Further developments regarding conservative sample size estimation under this constrained framework are reported in Section 7.

In our study it is assumed that 2 confirmative phase III trials are run simultaneously and independently. In accordance with [1] (first eq. of p.66), the success probability is that of finding two statistically significant results when M_n patients per group are recruited, giving $SP(M_n) = (\pi_{\delta_t}(M_n))^2$. Consequently, success probability is a random variable depending on the randomness of M_n and on its maximum $m_{\max}(r)$. The mean of $SP(M_n)$, conditional to $\mathcal{L}$, is of main interest and, although it has been called Average Power by Wang et al.[2], we call it the SP of phase III. Its formulation, in accordance with eq.3.10 in [1], is:

$$SP_{III}(r) = \sum_{m=2}^{m_{\max}(r)} SP(m) P_{\delta_t}(M_n = m|\mathcal{L}) \tag{3}$$

3 Defining OSP

Let us assume that the quantity to be optimized is the Overall Success Probability (OSP), that is the joined probability of success of phase II *and* phase III. In the recent past, OSP has been called Overall Power [8, 3, 1], however, we now tend to use the word *power* either to indicate power functions or to identify thresholds of desired success probability.

Since the (random) results of the two phases are independent - it is assumed that phase II data are not included in the analysis of phase III data, OSP is given by the product of the success probabilities of phase II and phase III. In particular, OSP is the probability of launching phase III *and* rejecting the null hypothesis during both phase III trials, when phase III sample size is estimated on the basis of phase II data and provided that the M_n results lower than $m_{\max}(r)$:

$$OSP(r) = SP_{II}(r) \times SP_{III}(r) = P_{\delta_t}(\mathcal{L}) \sum_{m=2}^{m_{\max}(r)} SP_{III}(m) P_{\delta_t}(M_n = m|\mathcal{L}) \tag{4}$$

which reduces to:

$$OSP(r) = \sum_{m=2}^{m_{\max}(r)} SP_{III}(m) P_{\delta_t}(M_n = m) \tag{5}$$

Moreover, from the exact distribution of d_n (i.e. $N(\delta_t, 2/n)$), that of M_n is derived and, in agreement to eq.3.10 in [1], we finally obtain:

$$OSP(r) = \sum_{m=2}^{m_{\max}(r)} [\Phi(\sqrt{\frac{m}{2}}\delta_t - z_{1-\alpha})]^2 \{\Phi(\sqrt{\frac{n}{2}}(\frac{z_{1-\alpha}+z_{1-\beta}}{\sqrt{(m-1)/2}} - \delta_t)) - \Phi(\sqrt{\frac{n}{2}}(\frac{z_{1-\alpha}+z_{1-\beta}}{\sqrt{m/2}} - \delta_t))\} \tag{6}$$

We expect $OSP(r)$ to be low for small values of r, due to low a launch probability (i.e. $SP_{II}(r)$), since the phase II sample size $n = rw/(h+1)$ is low. Also, $OSP(r)$ is expected to be low for high values of r, due to low values of $SP_{III}(M_n)$ since M_n is limited by a low value assumed by $m_{\max}(r)$ (through $w(1-r)/4$). In other words, when a small amount of resources is allocated to either phase II or phase III, the OSP should be low. Whereas, there exists an intermediate allocation of resources that optimizes $OSP(r)$. So, the problem is to find the optimal allocation $r_{opt} = argmax\{OSP(r)\}$.

4 Behavior of OSP

Some numerical examples are adopted to illustrate the behavior of OSP. At first, the settings are defined, that imply a simplification of launching rules. Then, the computing results are shown.

4.1 Settings

Three values of h, i.e. number of doses evaluated in phase II, are accounted for: $3, 5, 7$. As it concerns the launching rule, we start by focusing on the most intuitive, which is the one based on the threshold of clinical relevance δ_{0L}. It has recently been shown [3, 7], that δ_{0L} should be set not too close to δ_t, in order not to penalize SP_{II}. A threshold around $\delta_t/3$ is therefore set, accordingly. In phase III, the type I error probability is set at 2.5%, where the adopted power is $1 - \beta = 90\%$ - indeed, several papers in the field of sample size estimation suggest setting the power to at least the latter value (e.g. [2, 3]). Three effect size values are considered ($\delta_t = 0.2, 0.5, 0.8$), each one providing an ideal phase III sample size M_I which results $526, 85, 33$, respectively, from (1). For each δ_t, the whole sample size w is taken equal to kM_I, with five values of k: $10, 15, 20, 25, 30$. For each of the 45 settings (3 number of doses × 3 values of δ_t × 5 values of k), r is considered varying from 5% to 95% (step 1%).

4.2 Simplifying launching rule

Let us translate the launch threshold based on the effect size considered in the settings into that based on the maximum sample size, and then simplify the latter in order to ease OSP formulas. Having adopted the launch threshold of $\delta_t/3$, the maximum sample size allowed in phase III becomes $m_{\max} = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/(\delta_t/3)^2 \rfloor + 1$, i.e. approximately $9M_I$. For all the 45 settings defined above, the maximum sample size introduced by the constraint of available resources, i.e. $w(1 - r)/4$, results lower than $9M_I$. Consequently, $m_{\max}(r)$ turns out to be $w(1 - r)/4$. The OSP in equations (4) to (6) are, therefore, computed by replacing $m_{\max}(r)$ with $w(1 - r)/4$.

The actual launching threshold for the effect size, then, becomes $\delta_{0L}(r) = 2(z_{1-\alpha} + z_{1-\beta})\sqrt{2/w(1 - r)}$, since the latter emerges in all the settings higher than $\delta_t/3$. In other words, the launching rules based on the constraint on sample size given by the available resources (and that depend on r) are stricter than $\delta_t/3$. This is illustrated in Figure 1, where $\delta_{0L}(r)$ as a function of r is reported when $\delta_t = 0.5$. With the other values of δ_t the curves result very similar, as well as when the number of doses h varies from 3 to 7: in all the situation the thresholds are stricter than $\delta_t/3$. Hence, if $\delta_t/3$ is considered a threshold of clinical relevance, *a fortiori* $2(z_{1-\alpha} + z_{1-\beta})\sqrt{2/w(1 - r)}$ is so.

Note that this stricter launching rule, on one hand penalizes the probability of launching phase III, that is SP_{II}, but on the other is imposed by the model we are studying, which allocates w to phase II and phase III.

4.3 Computing OSP

The resulting values of $OSP(r)$, for different values of δ_t and k, but with $h = 5$ only, are reported in Figure 2. It is worth noting that the results under different δ_ts are very similar - they lie approximately on the same curves. Differences among $OSP(r)$ from different k levels are evident, and the values of $OSP(r)$ increase when k increases, i.e. when the whole sample size w increases.

When $k = 10$, $r_{opt} = 32\%$, regardless of δ_t; moreover $OSP(32\%) \simeq 44.5\%$, which is quite low with respect to the desired level of $SP_{III} = (90\%)^2 = 81\%$. Note also that when r is around 32%, the values of $OSP(r)$ are slightly lower: $OSP(20\%) \simeq 41\%$, and $OSP(50\%) \simeq 38\%$.

With $k = 20$, OSP is higher (see Figure 2), exceeding 70%. Once again, r_{opt}s under different δs are very close (i.e. $r_{opt} \simeq 46\%$), and we found $OSP(46\%) \simeq 71.7\%$: this is an interesting practical result, since 70% might look as an almost acceptable level of OSP. Moreover, when r is around 46%, the OSP is as follows: $OSP(25\%) \simeq 67\%$, and $OSP(65\%) \simeq 66\%$.

With $k \geq 25$, values of $OSP(r)$ are closer, especially when $r \in (30\%, 60\%)$. Moreover, r_{opt} moves from 52% (with $k = 25$) to 58% (with $k = 30$), meaning that if the whole set of resources increases (viz. k increases), the best solution is to allocate more and more sample size to phase II, in order to improve SP_{II} and the precision in estimating phase III sample size. Furthermore, with $k = 25$, $\max\{OSP\} = OSP(52\%) \simeq 76\%$, and with $k = 30$ we find $\max\{OSP\} = OSP(58\%) \simeq 79\%$: these findings suggest that when k increases, the maximum OSP tends to 81% (with $k = 100$, $\max\{OSP\} = 80.82\%$). In particular, this results is given by the fact that $SP_{II}(r_{opt})$ tends to 1 and $SP_{III}(r_{opt})$ tends to $(90\%)^2 = 81\%$.

In Figure 3, OSP curves with $h = 3$ to 7 doses are reported with $\delta_t = 0.5$ (black dots were already available in Figure 2): once again, $k \geq 20$ is suggested, even higher if $h = 7$. In these three dose settings, r_{opt} tends to be a bit higher than 50%, although $OSP(50\%)$ is almost equal to $OSP(r_{opt})$.

5 Sizing the whole amount of resources

The results of the above Section suggest that, when $h = 5$ doses are evaluated in phase II, at least k =20–25 should be adopted, that is, a whole sample size amounting to at least 20–25 times the ideal sample size of one group of phase III should be provided for phases II and III trials, regardless of the amplitude of the true effect size. It is not a big error to made an allocation which is not optimal: an error of $\pm 10\%$ (and in some cases of $\pm 20\%$) from the r_{opt} is well supported by the flexibility of the system (i.e. the pointwise sample size estimation framework). To provide insufficient overall resources (e.g. k =10–15) is much worse. In Table 1, the values of k providing a maximum OSP of at least 75% is given, with a number of phase II doses h from 1 to 9.

As a rule of thumb, in order to optimize the OSP of phases II and III, with a number of phase II groups ranging from 2 to 10 (and 2 phase III confirmatory trials) provide to the whole development project sufficient resources to recruit a number of patients from 20 to 30 times (increasing linearly with the number of groups) the ideal sample size of one treatment group of phase III, and allocate about 50% of the sample size to phase II, regardless of the amplitude of the effect size.

As regards the introductive example where h was 1, $k = 16$ is needed, and assuming an effect size of 0.5 which gives $M_I = 85$, resources for $1360 = 16 \times 85$ patients are indicated in order to achieve a 75% OSP. Then, $r_{opt} = 40\%$ of the sample size should be provided to phase II, that is $n = 272$ patients per group, and the remaining 816 patients can be enrolled, at most, in phase III, that is $M_{272} \leq 204$ (recall that 2 phase III trials are being considered now, i.e. 4 phase III groups). If the assumed effect size is right, the OSP is 75%.

5.1 Assuring the whole amount of resources

The problem that in practice δ_t is unknown does not influence the allocation choice based on OSP, since $OSP(r)$ is almost independent of δ_t, as has been shown. Nevertheless, to allocate enough resources to recruit, at most, a number of patients from 20 to 30 times M_I is required for about 75% of OSP,*where M_I depends on δ_t.*

So, in practice M_I should be replaced by $M_a = M(\delta_a) = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/\delta_a^2 \rfloor + 1$, where δ_a is the *assumed* effect size. Hence, we do not know how close M_a is to M_I. In order to reinforce the assumption on δ_a and limit parameter uncertainty, *assurance* can be applied [9]. This consists in defining a distribution around δ_a, call it $f_{\delta_a}(t)$, so that the assured sample size becomes $M_A = \int M(t)\, f_{\delta_a}(t)\, dt$. This technique can also be viewed as a Bayesian approach to sample size determination, where $f_{\delta_a}(t)$ plays the role of the prior distribution - indeed, recall that phase II data have not yet been observed.

Within Bayesian approaches, a Gaussian prior is often defined for the effect size. Here, we prefer to consider compact support prior distributions, such as uniform ones or the parabolic shaped ones of Epanechnikov kernel smoothers; this is because normal distributions allow for tail values of the effect size that may result too high or simply below zero, giving no solutions to the classical sample size formula. The prior distributions we tend to adopt are defined in a finite range around δ_a. When symmetric distributions are used, their density may be defined > 0 at most 50% above and below the central value δ_a, i.e. $f_{\delta_a}(t) > 0$ when $t \in (\delta_a(1 - \epsilon/2), \delta_a(1 + \epsilon/2)$, and $f_{\delta_a}(t) = 0$ otherwise, with $\epsilon \leq 1$, and $f_{\delta_a}(\delta_a - t) = f_{\delta_a}(\delta_a + t)$, $\forall t > 0$. For example, when the uniform prior is adopted with the maximum allowed variability, (i.e. $\epsilon = 1$, $f_{\delta_a}(t) = 1/(2\delta_a)$ when $t \in (\delta_a/2, 3\delta_a/2)$, and $f_{\delta_a}(t) = 0$ otherwise), it can easily be found that $M_A = 4M_a/3$: this is a situation where quite a high amount of assurance is applied, since the uniform shape assures more than the parabolic one and $\epsilon = 1$ allows a quite large variability around δ_a.

The rule of thumb above, through assurance, now suggests providing the whole development project when $h = 5$ with sufficient resources to recruit $24M_A$ patients, i.e. 32 times the assumed sample size M_a. Note that a lower assurance would provide $24 \leq k \leq 32$.

6 Mean and variability of total sample size

An indispensable aspect of this problem is the behavior of the sample size estimator M_n, reflecting the actual amount of resources spent in phase III and so those spent overall. M_n is a random variable that for small rs, i.e. when the phase II sample size n is low, might present large variability in estimating M_I: to compute the average of M_n, which is also related to the average cost of the trial, and its variability, the formulas are:

$$E[M_n|\mathcal{L}] = \sum_{m=2}^{m_{\max}(r)} m\, P_{\delta_t}(M_n = m|\mathcal{L})$$

and

$$MSE[M_n|\mathcal{L}] = \sum_{m=2}^{m_{\max}(r)} (m - M_I)^2\, P_{\delta_t}(M_n = m|\mathcal{L})$$

In Table 2, the average and the MSE of M_n, conditional to phase III launch, is shown for three different values of δ_t, each one implying a different M_Is, with k increasing from 15 to 30 (according to findings and suggestions of the previous Section), with $h = 5$ and with r equal to 25%, 50% and 75%. It can be noted that when the whole amount of resources increases (i.e. k increases) and the resource rate allocation to phase II (r) is fixed, both mean and MSE of $M_I|\mathcal{L}$ increase. Mainly, when r increases, the estimation process becomes more reliable: the mean of $M_I|\mathcal{L}$ tends to M_I and the MSE decreases, for each k of those considered.

Moreover, note that when $k = 25$ and $r = 50\%$ (that are the operating conditions giving high OSP when $h = 5$), the mean of M_n is close to M_I and the mean error is about $M_I/2$, for every δ_t. Indeed, the behavior of M_n too is almost independent of δ_t, not only that of OSP: Table 3 reports the standardized values of mean and mean error of Table 2, and it can be noted that the results of different δ_ts are very close.

Now, let us consider how these numbers reflect on the whole amount of resources spent in both phases, that is, on the total sample size: $M_T = M_I \times k \times r + 4M_n$, which is a random number too. From a practical standpoint, the settings with $k = 20, 25$

and $r = 25\%, 50\%$ are the most interesting. Indeed, when $k = 30$ a very high amount of resources is needed, and with $k = 15$ the OSP is often low; also, OSP is low with $r = 75\%$, due to strict constraints for M_n.

When $k = 25$ and $r = 50\%$, the average amount of resources actually spent is $E(M_T) = M_I \times (25 \times 50\% + 4 \times 1.11) \simeq 17M_I$, with a standard deviation of $\sigma(M_T) \simeq 2M_I$ - recall, this is almost independent of δ_t. It is also of great interest to obtain percentiles for M_n, and so for M_T: to compute them, the usual normal formula (i.e. average $+\Phi^{-1}(p) \times$ std.dev.) is not indicated since the distribution of M_n is not normal (e.g. Fig.3.2 in [1]); percentiles of M_n con be obtained through conditional probability calculation. For example, with $\delta_t = 0.5$ and under the latter setting where $n = 177$ and $E(M_{177}) = 94.0$, the 80% and 90% percentiles are $m_{177}^{.8} = 122$ and $m_{177}^{.9} = 151$, respectively. Once again, percentiles present small variations in function of δ_t.

Mean, standard deviation and percentiles of the four settings considered of main interest are reported in Table 4. In the light of these further results in terms of average sample size and variability, even rs that do not provide optimal OSP may be of practical interest. For example, when $k = 25$ is adopted (i.e., $w = 25M_I$ is allocated) and $r = 25\%$ of resources are used in phase II, the average of M_T is $11.1M_I$ and M_T does not overcome $12.8M_I$ with 80% probability, where $OSP = 72.4\%$.

7 Improving OSP through conservativeness

In the recent past, we underlined the importance of estimating the sample size conservatively [1]. This is in order to account for the variability of the effect size estimate and in agreement with some authoritative authors (e.g. [2, 5, 7]). In particular, we suggested considering a conservative estimator of δ_t, i.e. $d_n^\gamma = d_n - z_\gamma\sqrt{2/n}$, to be used in (2) in place of d_n. Some authors include this technique in adaptive ones [2], and its adoption "appears to be a reasonable choice" for planning phase III [10]. Moreover, sample size estimation performances can be improved by varying the amount of conservativeness γ [3].

In the problem here studied, the use of d_n^γ in OSP computation implies that (6) becomes $OSP_2(r, \gamma)$, which is a concave function of two variables - for the sake of brevity, we do not report the entire formulation here. Then, in practice OSP can be optimized for the amount of resources allocated to phase II *and* for the amount of conservativeness adopted in estimating phase III sample size. In Figure 4, $OSP_2(r, \gamma)$ is shown for one of the settings considered above, i.e. $\alpha = 2.5\%$, $1-\beta = 90\%$, $h = 5$, $\delta_t = 0.5$, and $k = 25$. The surface is concave and it presents a maximum at $(r, \gamma) = (46\%, 72\%)$, where $OSP_2(46\%, 72\%) = 80.1\%$ - remember that in this setting $OSP(r_{opt}) = 76.2\%$, and $r_{opt} = 52\%$.

In Table 5 some values of OSP_2 are reported, with the same settings of interest considered for building Table 1. It is worth noting that this approach may provide useful results when the possibility of exploiting conservativeness exists, that is when the constraint is not stringent. In particular, a fairly high amount of resources should be allocated to the whole development project, that is the choice of k in accordance to the second column of Table 1. Under these conditions, OSP_2 can be increased by about 3%: this improvement seems low, whereas it might actually be a lot when multiplied by the revenues produced if the trial succeeds.

To conclude, consider the setting of possible practical interest, although not optimal, discussed in last Section, that was $k = 25$ and $r = 25\%$: with $\gamma = 68\%$, the average and the 80%-tile of the total sample size M_T increase to $11.8M_I$, and $14.9M_I$, respectively, where OSP becomes 75.6%, i.e. $+3.2\%$.

8 Discussion

Although the development of a drug, and in particular the clinical part regarding phase II and III trials, might be looked at in its entirety, scientists and trial managers often tend to focus on each phase separately. It is a fact that the failure rate of phase II and phase III clinical trials is quite high. As a consequence, the overall success provability (OSP) has recently begun to be used as a tool for planning experiments [8, 4, 3]. In this paper, we studied the problem of allocating the available resources for developing phase II and III trials (in terms of sample size) to each of the two phases. It was assumed that 2 phase III trials are run with a sample size estimated on the basis of phase II data. Overall success probability has been evaluated, and the variability of the resources actually spent has been accounted for.

We showed that to obtain a high OSP (e.g. 75%) the whole amount of resources needed is one order of magnitude higher than the ideal sample size of one group of one phase III trial (viz. M_I). In particular, when the number of doses evaluated in phase II goes from 3 to 9, the whole amount of resources needed varies from 20 to 31 times M_I. This is almost independent of the effect size of the dose selected in phase II. Moreover, to obtain the optimal OSP, the rate of resources to be allocated to phase II is often close to 50%. It should be remarked that the remaining resources allocated to phase III are not necessarily spent, they are just available, and their use depends on how large the phase III sample size estimated through phase II data is. The higher the rate allocated to phase II, the more precise sample size estimation is, provided that enough resources are left for phase III. Even an amount of resources as low as 25% might give an acceptable OSP and an invitingly small total sample size if allocated to phase II, provided that a sufficient amount of resources is allocated to the two phases.

Concerning phase III sample size estimation, conservativeness may be adopted (see [1]), and may result in a considerable increase in OSP (i.e. about 3% when $OSP \simeq 75\%$). If the whole amount of resources available for the two phases is low (i.e. lower than 15 times M_I), the OSP will be low too, even lower than 50%, even if the best allocation of resources is made between the two phases together with conservative sample size estimation. Since M_I depends on the unknown effect size of the selected dose, wrong assumptions regarding the latter can cause too small investments and low OSP. To reduce this risk, M_I my be computed by applying assurance [9] on effect size assumptions.

The indications on the amount of resources to be allocated to phase II suggested by Jiang [4] differ from ours, but in that paper only 2 phase II groups and 1 phase III trial are taken into account. Differences between our indications and those provided by Stallard [11] are much more evident, since phase II data are considered there only for detecting a certain effect with low power, not for adequately planning phase III. Often, papers in the field of sample size estimation adopt the Bayesian approach: although in simple situations Bayesian estimators present a very high variability and are, therefore, not indicated for practical purposes [3], in more complex ones (e.g. [12]) they can be of some interest, especially is appropriated launching rules on the maximum allowed sample size are adopted jointly.

In the last decade, there has been an increase in the adoption of adaptive designs in phase II and phase III, but fixed sample designs are still used in the great majority of clinical trials. Wang et al. [10] suggest the adoption of adaptive design in exploratory trials in order to reduce costs and durations of the studies, whenever the more complicated logistics and a correct methodology allows their application. In this context, instead of stopping rules based on statistical significance or on the precision of the effect size estimate, in accordance with the exploratory aim of phase II trials we suggest the adoption of stopping rules based on the precision of the sample size estimator for phase III (i.e. M_n), provided that clinically meaningful results are observed.

The 50% allocation suggested here in order to optimize OSP is usually not adopted in clinical practice: phase II often absorbs less resources than phase III. Indeed, the size of samples adopted in phase II is, on average, 10–15% of the total sample size of the two phases (where 2 phase III trials are considered) [1]. To improve the success rate of phase II and phase III trials, phase II allocation might be increased to, at least, 25%, provided that a sufficient global amount of resources is available. Then, a more accurate phase II would also induce a higher probability of choosing the best dose among those considered. Nevertheless, larger phase II trials imply higher costs and longer times for the development project, allowing for a shorter patent life and so lower potential gains, of course in case of successful trials. Optimal allocation of resources should also be evaluated from an economic perspective, as suggested by Jiang [4] too. For this reason, our future works may focus on the relationship between allocations, OSP, efficacy and safety utility functions, costs, revenues, and gain, according to [13, 14].

References

1. De Martini D. *Success Probability Estimation with Applications to Clinical Trials.* John Wiley & Sons: Hoboken, NJ, 2013.

2. Wang SJ, Hung HMJ, O'Neill RT. Adapting the sample size planning of a phase III trial based on phase II data. *Pharmaceutical Statistics* 2006; **5**: 85–97.

3. De Martini D. Adapting by calibration the sample size of a phase III trial on the basis of phase II data. *Pharmaceutical Statistics* 2011; **10** (2): 89–95.

4. Jiang K. Optimal Sample Sizes and Go/No-Go Decisions for Phase II/III Development Programs Based on Probability of Success. *Statistics in Biopharmaceutical Rerearch* 2011; **3**: 463-475.

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

6. De Martini D. Robustness and corrections for sample size adaptation strategies based on effect size estimation. *Communications in Statistics - Simulation and Computation* 2011; **40** (9): 1263–1277.

7. Kirby S, Burke J, Chuang-Stein C, Sin C. Discounting phase 2 results when planning phase 3 clinical trials. *Pharmaceutical Statistics* 2012; **11**, 5: 373–385.

8. Fay MP, Halloran ME, Follmann DA. Accounting for Variability in Sample Size Estimation with Applications to Nonhaderence and Estimation of Variance and Effect Size. *Biometrics* 2007; **63**: 465–474.

9. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharmaceutical Statistics* 2005; **4**: 187–201.

10. Wang SJ, Hung HMJ, O'Neill RT. Paradigm for adaptive statistical information designs: practical experiences and strategies. *Statistics in Medicine* 2012; **31**: 3011–3023.

11. Stallard N. Optimal sample sizes for phase II clinical trials and pilot studies. *Statistics in Medicine* 2012; **31**: 1031–1042.

12. Wang Y, Fu H, Kulkarni P, Kaiser C. Evaluating and utilizing probability of study success in clinical development. *Clinical Trials* 2013; **0**; 1–7.

13. Patel N., Bolognese J., Chuang-Stein C., Hewitt D., Gammaitoni A., Pinheiro J. Designing Phase 2 Trials Based on Program-Level Considerations: A Case Study for Neuropathic Pain. *Drug Information Journal* 2012; **46**, 4: 439–454.

14. Chen MH., Willian AR. Determining optimal sample size for multistage adaptive randomized clinical trials from an industry perspective using value of information methods. *Clinical Trails* 2013; **10**: 54–62.

Table 1	Values of k to join 75% of OSP		
h	k	r_{opt}	$OSP(r_{opt})$
1	16	40%	76.2%
2	17	43%	75.3%
3	20	47%	75.9%
4	22	51%	75.7%
5	24	52%	75.6%
6	26	53%	75.6%
7	27	53%	75.1%
8	29	56%	75.2%
9	31	56%	75.3%

Table 1. Minimum values of k to join a max OSP of at least 75%, with $\alpha = 0.025$, $1-\beta = 0.9$, $h = 1, \ldots, 9$, and $\delta_t = 0.5$.

Table 2	Mean and MSE of SSE						
δ_t	k	$r = 25\%$		$r = 50\%$		$r = 75\%$	
		$E[M_n\|\mathcal{L}]$	$MSE[M_n\|\mathcal{L}]$	$E[M_n\|\mathcal{L}]$	$MSE[M_n\|\mathcal{L}]$	$E[M_n\|\mathcal{L}]$	$MSE[M_n\|\mathcal{L}]$
0.2	15	542.3	95643.6	504.6	38078.7	376.5	27581.7
$(M_I = 526)$	20	606.9	145686.2	561.9	58211.7	458.0	15703.9
	25	638.1	181566.1	584.4	70551.9	512.4	18439.4
	30	647.8	195765.9	587.8	71008.5	541.4	24290.4
0.5	15	87.6	2482.1	81.4	986.8	61.3	703.5
$(M_I = 85)$	20	97.9	3800.9	90.6	1510.3	74.4	410.2
	25	102.7	4682.1	94.0	1799.1	82.7	481.8
	30	104.1	5006.8	94.5	1809.1	87.3	631.4
0.8	15	34.3	379.9	32.1	149.5	24.5	95.0
$(M_I = 33)$	20	38.4	580.5	35.6	227.5	29.4	59.2
	25	40.4	721.4	37.1	283.2	32.4	70.2
	30	41.0	775.6	37.2	282.1	34.4	99.4

Table 2. Mean and MSE of M_n, with $\alpha = 0.025$, $1 - \beta = 0.9$, $h = 5$, $\delta_t = 0.2, 0.5, 0.8$, $k = 15, 20, 25, 30$ and $r = 25\%, 50\%, 75\%$.

Table 3	Standardized Mean and MSE of SSE						
δ_t	k	$r = 25\%$		$r = 50\%$		$r = 75\%$	
		$E[\bullet]/M_I$	$\sqrt{MSE[\bullet]}/M_I$	$E[\bullet]/M_I$	$\sqrt{MSE[\bullet]}/M_I$	$E[\bullet]/M_I$	$\sqrt{MSE[\bullet]}/M_I$
0.2	15	1.03	0.588	0.96	0.371	0.72	0.316
	20	1.15	0.726	1.07	0.459	0.87	0.238
	25	1.21	0.810	1.11	0.505	0.97	0.258
	30	1.23	0.841	1.12	0.507	1.03	0.296
0.5	15	1.03	0.586	0.96	0.370	0.72	0.312
	20	1.15	0.725	1.07	0.457	0.88	0.238
	25	1.21	0.805	1.11	0.499	0.97	0.258
	30	1.22	0.832	1.11	0.500	1.03	0.296
0.8	15	1.04	0.591	0.97	0.371	0.74	0.295
	20	1.16	0.730	1.08	0.457	0.89	0.233
	25	1.22	0.814	1.12	0.510	0.98	0.254
	30	1.24	0.844	1.13	0.509	1.04	0.302

Table 3. Standardized mean and MSE of M_n, with $\alpha = 0.025,\ 1-\beta = 0.9,\ h = 5,\ \delta_t = 0.2, 0.5, 0.8,\ k = 15, 20, 25, 30$ and $r = 25\%, 50\%, 75\%$ (the symbol "$\bullet$" stands for "$M_n|\mathcal{L}$").

Table 4	Standardized measures of total expenses							
k	r	$E(M_T)$	$\sigma(M_T)$	$m_n^{.8}$	$m_n^{.9}$	$m_T^{.8}$	$m_T^{.9}$	OSP
20	$r = 25\%$	$9.6M_I$	$2.8M_I$	$1.6M_I$	$2.2M_I$	$11.5M_I$	$13.8M_I$	67.7%
	$r = 50\%$	$14.3M_I$	$1.8M_I$	$1.5M_I$	$1.7M_I$	$16.0M_I$	$16.9M_I$	71.7%
25	$r = 25\%$	$11.1M_I$	$3.1M_I$	$1.6M_I$	$2.2M_I$	$12.8M_I$	$15.2M_I$	72.4%
	$r = 50\%$	$16.9M_I$	$2.0M_I$	$1.4M_I$	$1.8M_I$	$18.2M_I$	$19.6M_I$	76.2%

Table 4. Standardized mean, st.dev. and percentiles of the total expenses in terms of sample size (viz. M_T), through percentiles of M_n, obtained with $\alpha = 0.025$, $1-\beta = 0.9$, $h = 5$, $k = 20, 25$ and $r = 25\%, 50\%$; also, $\delta_t = 0.5$ has been adopted.

Table 5	Maximum values of OSP_2			
h	k	r_{opt}	γ_{opt}	$OSP_2(r_{opt}, \gamma_{opt})$
3	20	42%	69%	78.7%
5	24	46%	70%	78.9%
7	27	48%	70%	78.3%
9	31	50%	72%	79.2%

Table 5. Values of OSP_2 for same remarkable settings and values of k from Table 1.

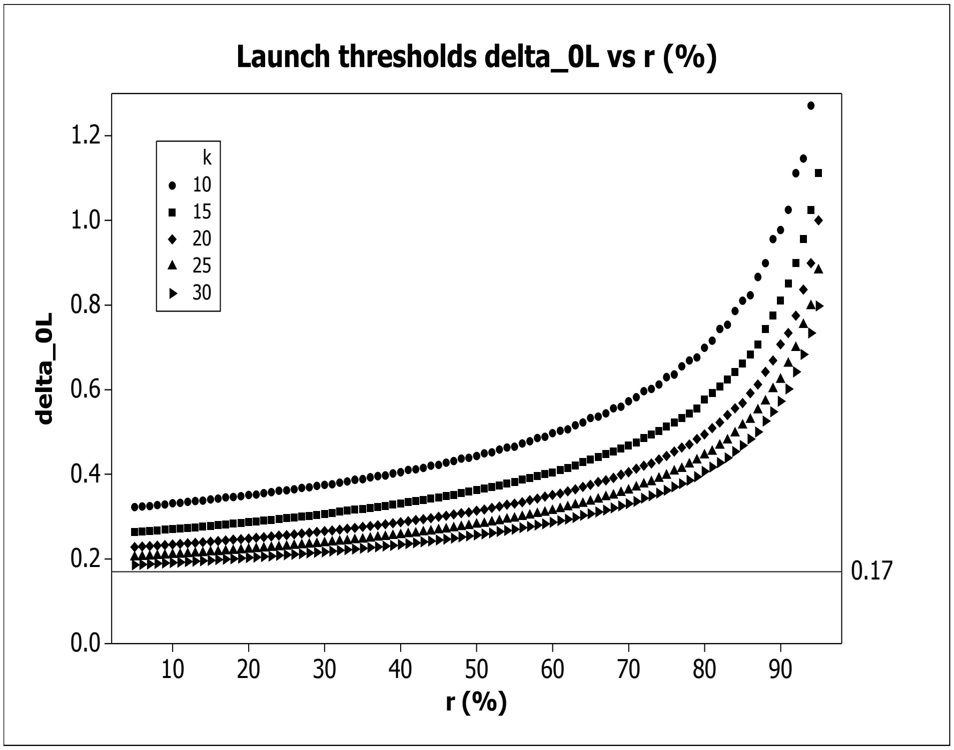

Figure 1: *Launch thresholds* δ_{0L}*, obtained with* $\alpha = 2.5\%$*,* $1 - \beta = 90\%$*,* $\delta_t = 0.5$*, and with* $k = 10, 15, 20, 25, 30$.

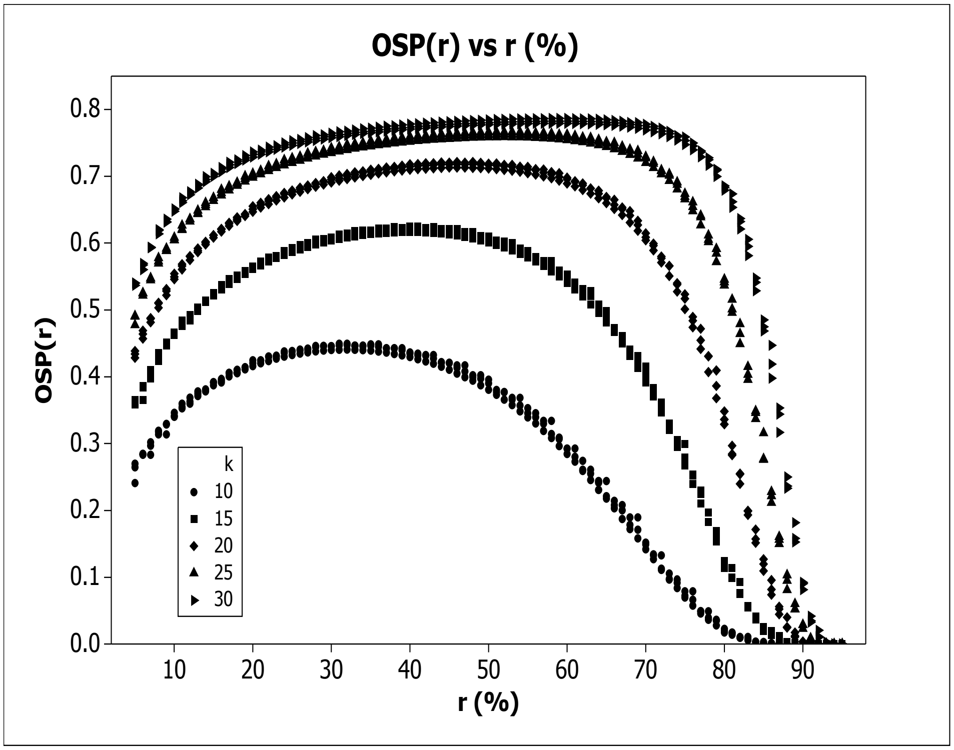

Figure 2: *$OSP(r)$, obtained with $\alpha = 2.5\%$, $1 - \beta = 90\%$, $h = 5$, $\delta_t = 0.2, 0.5, 0.8$, and with $k = 10, 15, 20, 25, 30$.*

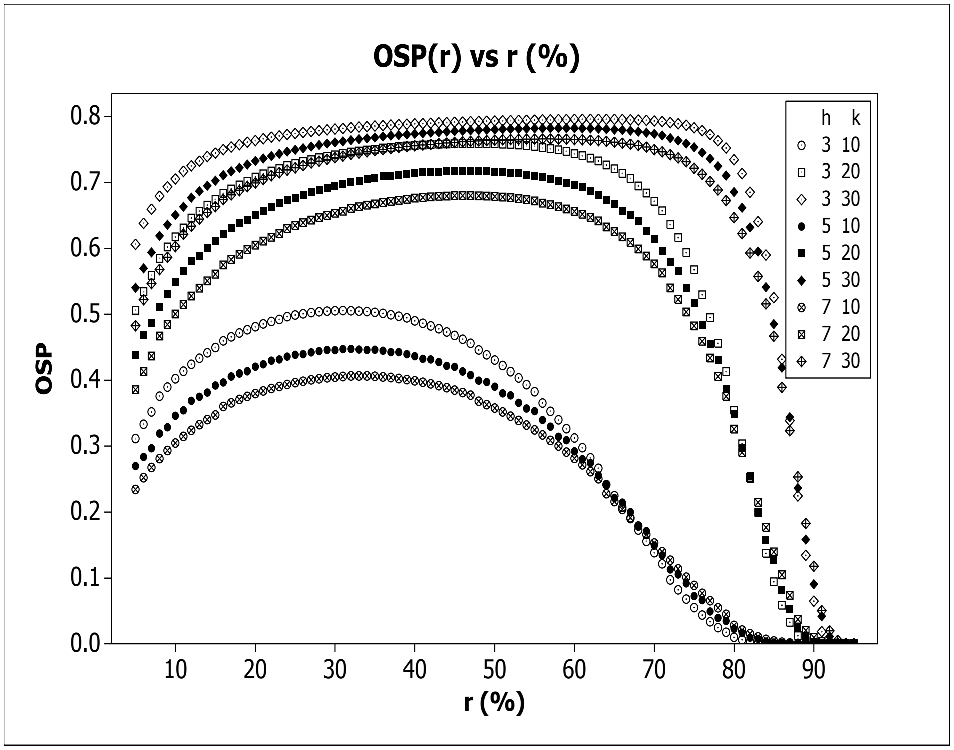

Figure 3: *$OSP(r)$, obtained with $\alpha = 2.5\%$, $1 - \beta = 90\%$, $h = 3, 5, 7$, $\delta_t = 0.5$, and with $k = 10, 20, 30$.*

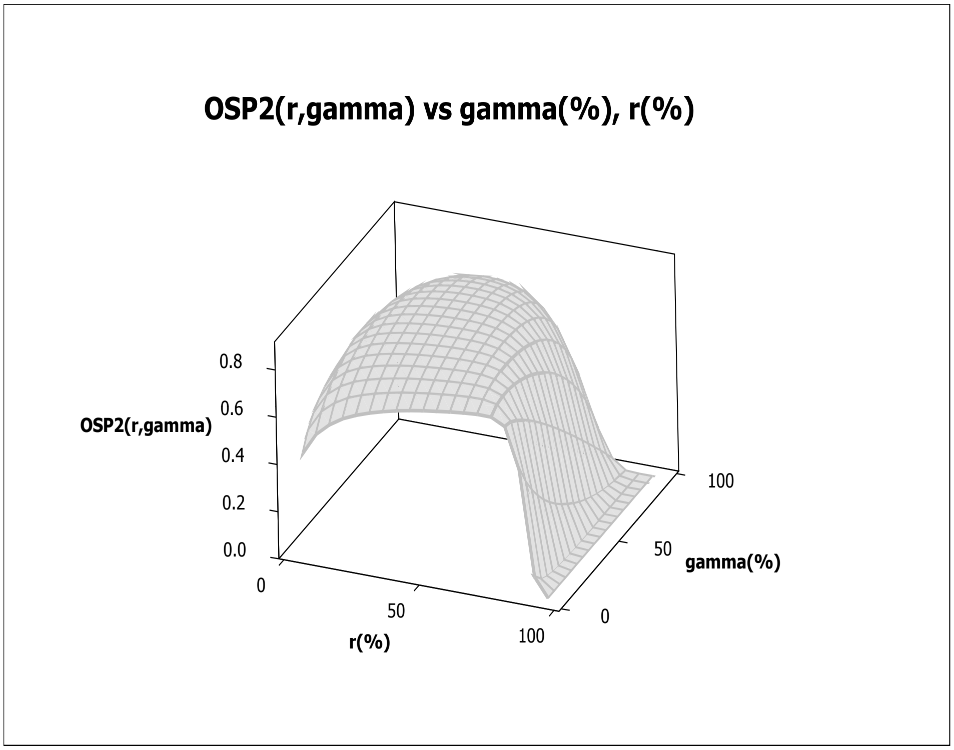

Figure 4: $OSP_2(r, \gamma)$, *obtained with* $\alpha = 2.5\%$, $1 - \beta = 90\%$, $h = 5$, $\delta_t = 0.5$, *and with* $k = 25$.