Allocating the Sample Size in Phase II and III Trials to Optimize Success Probability (Resource Allocation to Optimize Success)

Daniele De Martini[1]

ABSTRACT

BACKGROUND: 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 to increase the success rate. The overall success probability (OSP) is accounted for.

METHODS: Focus is placed on the amount of resources that should be provided to phase II and III trials to attain a good level of OSP, and on how many of these resources should be allocated to phase II to optimize OSP. It is assumed that phase II data are not considered for conrmatory purposes and are used for planning phase III through sample size estimation. Being r the rate of resources allocated to phase II, $OSP(r)$ is a concave function and there is an optimal allocation r_{opt} giving $max(OSP)$. If M_1 is the sample size giving the desired power to phase III, and kM_1 is the whole sample size that can be allocated to the two phases, how large k and r should be is indicated in order to achieve levels of OSP of practical interest.

RESULTS: For example, when 5 doses are evaluated in phase II and 2 parallel phase III conrmatory 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%. The choice of k depends mainly on how many phase II treatment groups are considered, not on the eect size of the selected dose. When k is large enough, r_{opt} is close to 50%. An $r \simeq$ 25%, although not best, might give a good OSP and an invitingly small total sample size, provided that k is large enough.

CONCLUSIONS: To improve the success rate of phase II and phase III trials, the drug development could be looked at in its entirety. Resources larger than those usually employed should be allocated to phase II to increase OSP. Phase II allocation rate may be increased to, at least, 25%, provided that a sucient global amount of resources is available.

Key words: overall success probability; allocation rate; launching rules; sample size estimation; optimal allocation; sucient resources.

(1) Dipartimento DiSMeQ - Universita degli Studi di Milano-Bicocca

CORRESPONDING AUTHOR: *Daniele De Martini- Dipartimento DiSMeQ - Universita degli Studi di Milano-Bicocca - Via Bicocca degli Arcimboldi 8, 20126 Milano - Italia - E-mail: daniele.demartini@unimib.it Tel.+39-02-6448-3130*

DOI: ***10.2427/9958***

Accepted on 22 July, 2014; Published as Online First on 15 December, 2014.

1 Introduction

It is common knowledge that the aim of phase II clinical trials 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, and the rate of trial failures, which is around 60% and 40% for phase II and III respectively suggests that this habit might not be helpful. In general, low success probabilities are often due to low sample size [1]. 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 has been run, with 2 parallel arms each with 60 patients, and that a phase III with the same design needs to be planned. Assume that the efficacy value (standardized effect size) of minimum interest that should be observed to then launch phase III is 0.15 and that a value slightly higher than this has been observed in phase II, so that phase III has to be launched. With one-sided $\alpha = 2.5\%$ and power $1 - \beta = 90\%$, approximately 940 patients should be recruited for each group, if the observed effect size is adopted for sample size computation - namely pointwise strategy [2, 3]. This number (940) is quite high, but not beyond the range of those usually adopted in phase III trials (visit *clinicaltrials.gov*, a service of the U.S. NIH). So, assuming that the research team decided to actually launch phase III, the total number of patients enrolled in about 2,000.

Now, the point is: if the resources for studying 2,000 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 when 400 data per group come from phase II and are used for estimating the phase III sample size, this 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, the aims of phase II are to correctly decide go/no-go, to launch phase III (i.e. go) with a high probability if a meaningful effect really exists, and to estimate well 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.
In this paper we study sample size resource allocation in terms of overall probability of success (OSP): 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 many of these resources should be allocated to phase II to optimize OSP. Since not all the resources allocated to phase III are spent, depending on sample size estimation, we also focus on the actual amount of resources used. 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 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].

2 Theoretical framework

2.1 Drug development model

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 for 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.
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. Let $r \in (0, 1)$ be the rate of w allocated to phase II: if a sample of size n is studied for each treatment in phase II, then n 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 (M_n) is estimated on the basis of phase II data and is a random variable whose maximum is $w(1 - r)/4$ (4 are phase III groups).

2.2 Phase II tools

Let $\delta = (\mu_D - \mu_P)/\sigma$ be the generic standardized effect size, μ_D and μ_P the means of response variables of the populations under D and under placebo. Without loss of generality $\sigma = 1$. The true, unknown, effect size is δ_t. $\bar{X}_{D,n}$ and $\bar{X}_{P,n}$ are the means of measurements and $d_n = \bar{X}_{D,n} - \bar{X}_{P,n}$ is the pointwise estimator of δ_t.
Call $\mathcal{L}$ the random event representing the success of phase II - $\mathcal{L}$ stands for phase III *launch*. $\mathcal{L}$ can be defined in some different ways: for example, on the basis of the maximum sample size $m_{\max}$ for phase III (i.e. $\mathcal{L} \Leftrightarrow M_n \leq m_{\max}$), or the basis of the observed effect size overcoming a threshold of clinical relevance (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}. Note that the first two launching criteria above can be set to result mathematically equivalent ([1], Ch.3).
Although the launching rule based on δ_{0L} is the most intuitive, and one of the most used, let us adopt the one pragmatically based on $m_{\max}$. In this framework constrained by w and modeled by r, the actual launching rule becomes $M_n \leq m_{\max}(r) = \min\{m_{\max}, w(1 - r)/4\}$. 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 3.2).

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 Z-test is adopted with one-sided alternatives. Being m the generic sample size, $T_m = \sqrt{m/2}(\bar{X}_{D,m} - \bar{X}_{P,m})$ is the test statistic and the success probability, according to [1], Ch.3, is $P_{\delta_t}(T_m > z_{1-\alpha}) = SP(m)$.
$1-\beta$ is the desired power to be achieved in each phase III trial (e.g. 90%); then, the ideal sample size per group for each phase III trial is:

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

Once phase II has succeeded, phase III is run with the sample size estimated by the $2n$ phase II data. Several sample size estimation strategies can be adopted [1]. Here, M_I is estimated by the pointwise estimator based on the observed effect size 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 its performances in terms of OSP and MSE are acceptable although not best [3].
Two confirmative phase III trials are run simultaneously and independently, each group with M_n patients, so that the success probability is the random variable $(SP(M_n))^2$. The mean of $(SP(M_n))^2$, 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:

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

2.4 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]). Since the 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 phases II and III:

$$OSP(r) = SP_{II}(r) \times SP_{III}(r) = \sum_{m=2}^{m_{\max}(r)} (SP(m))^2 P_{\delta_t}(M_n = m) \quad (4)$$

We expect $OSP(r)$ to be low for small values of r, due to a low launch probability (i.e. $SP_{II}(r)$). 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 $m_{\max}(r)$. Then, there is an intermediate allocation of resources that optimizes $OSP(r)$, that is $r_{opt} = argmax\{OSP(r)\}$.

3 Behavior of OSP

3.1 Settings

It has recently been shown [3, 7], that the threshold of clinical relevance 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. Phase III type I error is 2.5%, where the power is $1-\beta = 90\%$ (according to [2, 3]). Three effect size values are considered ($\delta_t = 0.2, 0.5, 0.8$), providing ideal phase III sample size M_Is resulting $526, 85, 33$, from (**??**). For each δ_t, the whole sample size w is taken equal to kM_I, with $k = 10, 15, 20, 25, 30$. Three numbers of doses h are accounted for: $3, 5, 7$. For each of the 45 settings (3 δ_ts $\times$ 5 ks $\times$ 3 hs), r is considered varying from 5% to 95%.

3.2 Simplifying launching rule

Let us translate the launch threshold based on the effect size into that based on the maximum sample size, and then simplify OSP formulas. Being $\delta_{0L} = \delta_t/3$, we have $m_{\max} = \lfloor 2(z_{1-\alpha} + z_{1-\beta})^2/(\delta_t/3)^2 \rfloor + 1 \simeq 9M_I$. For all 45 settings, 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 (**??**) is computed by replacing $m_{\max}(r)$ with $w(1-r)/4$. The actual launching threshold for the effect size 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 $\delta_t/3$: the launching rules

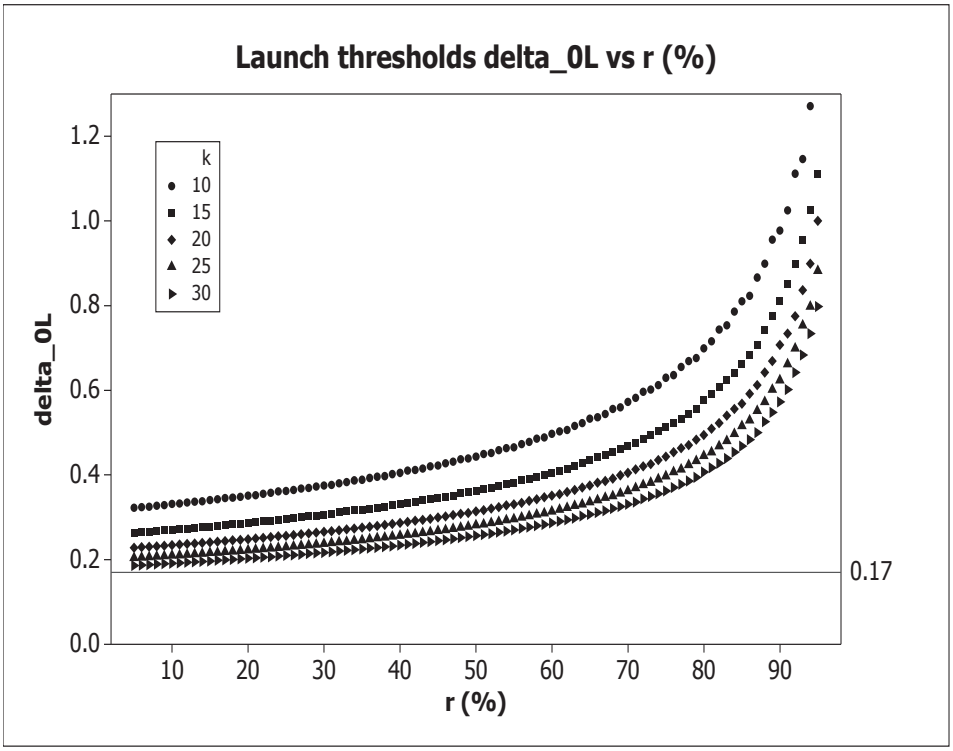

Figure 1: *Launch thresholds* δ_{0L}*, obtained with* $\alpha = 2.5\%$*,* $1 - \beta = 90\%$*,* $\delta_t = 0.5$*, and with* $k = 10, 15, 20, 25, 30$.

based on the constraint on sample size given by the available resources are stricter than $\delta_t/3$ (see Figure **??**, where $\delta_{0L}(r)$ as a function of r is reported - varying δ_ts and hs the curves result very similar). 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 penalizing the probability of launching phase III (i.e. SP_{II}) is imposed by the model we are studying.

3.3 Computing OSP

OSP functions, with $h = 5$, are reported in Figure **??**: $OSP(r)$ under different δ_ts are very similar - they lie approximately on the same curves.
Differences among $OSP(r)$ from different ks are evident: the values of $OSP(r)$ increase when k increases. OSP levels that look acceptable are obtained when k is

at least 20 (see Figure ??). When $k = 10, 15$, we have $\max\{OSP\} = OSP(r_{opt}) \simeq 45\%$, and $\max\{OSP\} \simeq 63\%$, respectively. With $k = 20$, $\max\{OSP\} \simeq 72\%$ - the optimal rates r_{opt}s under different δs are very close (i.e. $r_{opt} \simeq 47\%$). For $k = 25, 30$, we have $\max\{OSP\} \simeq 76\%, 78\%$, with $r_{opt} \simeq 52\%, 60\%$, respectively.

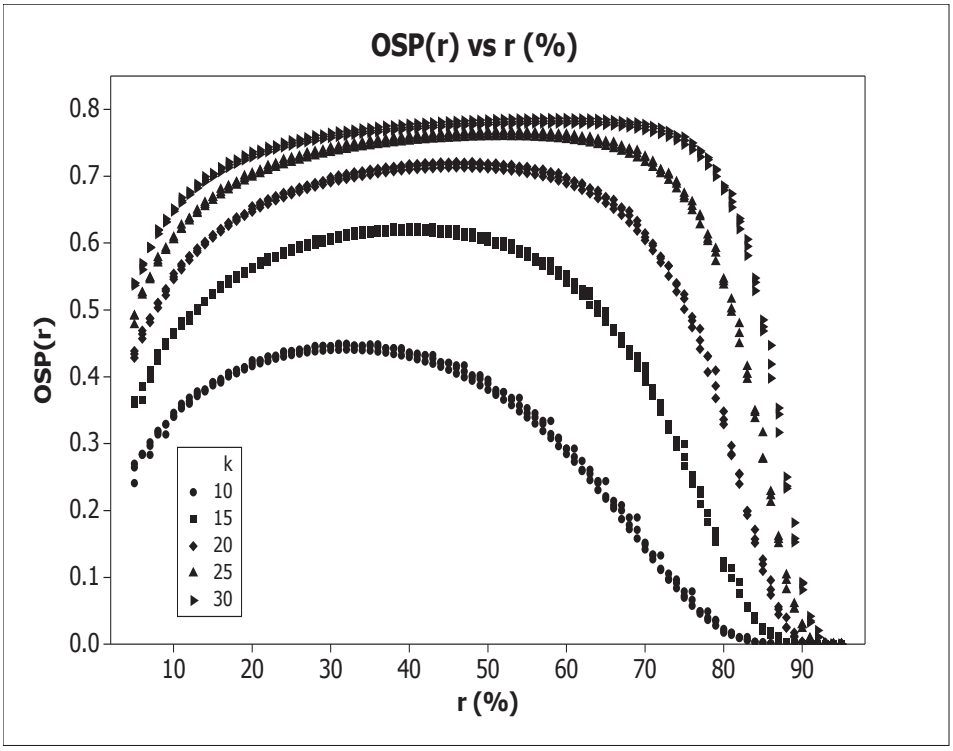

Figure 2: *$OSP(r)$, obtained with $\alpha = 2.5\%$, $1 - \beta = 90\%$, $h = 5$, $\delta_t = 0.2, 0.5, 0.8$, and with $k = 10, 15, 20, 25, 30$.*

For $k \geq 20$, $OSP(r)$ shows a quite flat shape around its maximum r_{opt}, meaning that even if the rate r allocated to phase II is a bit smaller than r_{opt}, $OSP(r)$ still provides acceptable values. For example, $OSP(30\%) \simeq 70\%, 74\%, 76\%$, with $k = 20, 25, 30$, respectively.

Finally, r_{opt} moves from 34% to 60% with k increasing from 10 to 30: when w increases, the best solution is to allocate more and more sample size to phase II, to improve both SP_{II} and the precision in estimating M_I.

In Figure ??, OSP curves with $h = 3, 5, 7$ are reported ($\delta_t = 0.5$): k set around 20

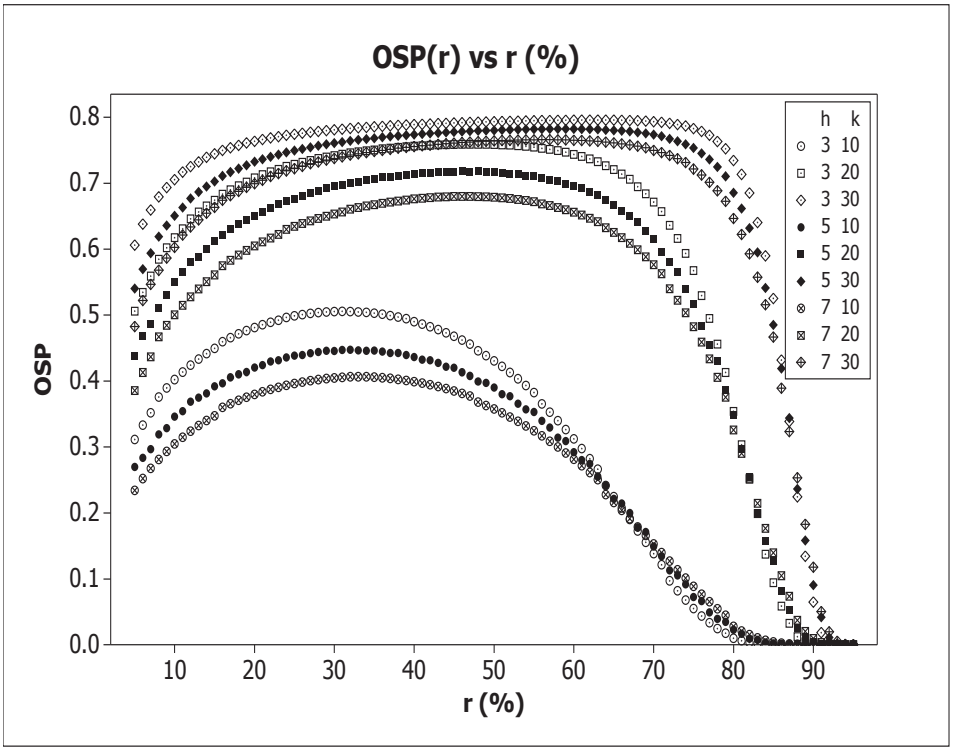

Figure 3: *OSP*(r), *obtained with* $\alpha = 2.5\%$, $1-\beta = 90\%$, $h = 3,5,7$, $\delta_t = 0.5$, *and with* $k = 10,20,30$.

(even higher if $h = 7$) is suggested to reach suitable OSP values.

4 Sizing the whole amount of resources

Since the desired SP_{III} is $(90\%)^2 = 81\%$, and a good SP_{II} level may be still around 90% - we remind that the aim is to study how to increase the success rates in clinical trials, we consider OSP around 72% as acceptable.
In Table 1, the values of k providing a maximum OSP of at least $70\%, 75\%$ (viz. $k_{70\%}, k_{75\%}$), are given, together with r_{opt}, with h from 1 to 9. Moreover, when $k_{75\%}$s are computed, also the smallest r giving an OSP at least 70% is given (viz. $r_{70\%} = \min\{r\, s.t.\, OSP(r) \geq 70\%\}$).

The total amount of resources needed to reach $OSP = 70\%$ is, with $h = 3, 5, 7$, $w = 16M_I, 19M_I, 22M_I$, respectively, with $r_{opt} = 40\%, 45\%, 48\%$. If smaller rs are allocated to phase II, OSP falls below 70% (e.g. $OSP(20\%) \simeq 64\%$).
When the 75% level is adopted, higher ks are needed: with $h = 3, 5, 7$, $k = 19, 24, 26$, respectively, provide the required OSP with $r_{opt} = 45\%, 50\%, 53\%$. However, if smaller rs are adopted, acceptable OSP levels are still provided: rs around 20% give $OSP(r) \simeq 70\%$.

Table 1	Values of k to join 70% and 75% of OSP				
h	$k_{70\%}$	r_{opt}	$k_{75\%}$	r_{opt}	$r_{70\%}$
1	13	35%	15	40%	15%
2	15	40%	17	42%	19%
3	16	40%	19	45%	20%
4	18	43%	22	49%	21%
5	19	45%	24	50%	21%
6	20	47%	25	51%	24%
7	22	48%	26	53%	24%
8	23	48%	29	56%	24%
9	24	50%	31	57%	24%

Table 1. Minimum values of k to join a max OSP of at least 70%, 75% (viz. $k_{70\%}, k_{75\%}$), with $\alpha = 0.025$, $1 - \beta = 0.9$, $h = 1, \ldots, 9$, and $\delta_t = 0.5$, together with the allocations r_{opt} providing max of OSP. Moreover, $r_{70\%}$ indicates the smallest allocation providing OSP at least 70%, when $k_{75\%}$ is adopted.

As a rule of thumb, in order to obtain an $OSP \simeq 75\%$, 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 15 to 30 times (increasing linearly with h) the ideal sample size M_I, and allocate about 50% of the sample size to phase II, regardless of the amplitude of δ_t.
Moreover, if just 20% of resources is allocated to phase II, OSP remains near 70%, provided that resources to reach $OSP = 75\%$ are stored before starting phase II.
We remark that not all the stored resources are used: rw is actually spent in phase II, where the global phase III sample size $4M_n$ is at most $(1 - r)w$ (see next Section).

4.1 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. Nevertheless, to allocate enough resources to obtain a given OSP level, depends on δ_t. In particular, since we adopted M_I as a unit measure for w, the resources needed depends on δ_t through M_I.

In practice, the unknown 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. However, how close M_a is to M_I is unknown. To reinforce the assumption on δ_a and limit the uncertainty of parameter, *assurance* can be applied [9]. This consists in defining a distribution around δ_a (viz. $f_{\delta_a}(t)$) so that the assured sample size becomes $M_A = \int M(t)\, f_{\delta_a}(t)\, dt$ - it can be viewed as Bayesian sample size determination, where $f_{\delta_a}(t)$ plays the role of the prior distribution.

For example, when the uniform prior $f_{\delta_a}(t) = 1/(2\delta_a)$, $t \in (\delta_a/2, 3\delta_a/2)$, is adopted, we find $M_A = 4M_a/3$.

The linear rule of thumb above, through assurance, suggests providing the whole development project when $h = 5$ with sufficient resources to recruit $22.5 M_A$ patients, i.e. $22.5 \times 4/3 = 30$ times the assumed sample size M_a. A lower assurance provides $22.5 \leq k \leq 30$.

5 Mean and variability of total sample size

An indispensable aspect of this sample size allocation problem is to evaluate the actual amount of resources spent in phase III, as well as those spent overall, depending on the behavior of the sample size estimator M_n.

M_n is a random variable that for small rs, i.e. when n is low, might be imprecise: the average and the MSE (measuring variability) of M_n, conditional to phase III launch, are:

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

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

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 $r = 25\%, 50\%, 75\%$.

In Table 2, the average and the MSE are shown with $k = 15, 20, 25, 30$, $h = 5$, and with $r = 25\%, 50\%, 75\%$. When k increases and r is fixed, both mean and MSE of $M_n|\mathcal{L}$ increase. Mainly, the estimation process becomes more reliable when r increases: the mean of $M_n|\mathcal{L}$ tends to M_I and MSE decreases.
Moreover, when $k = 25$ and $r = 50\%$ (viz. 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 is almost independent of δ_t, in accordance with that of OSP.
Now, let us consider how these numbers reflect on the whole amount of resources spent in both phases, viz. on the total sample size $M_T = M_I \times k \times r + 4M_n$. From a practical standpoint, the settings with $k = 20, 25$ and $r = 25\%, 50\%$ are the most interesting - $k = 30$ provides OSP higher than requested, 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 $r = 50\%$, and with $\delta_t = 0.5$ giving $M_I = 85$, the average amount of resources spent is $E(M_T) = 85 \times 25 \times 50\% + 4 \times 94.0 = 1438.5 \simeq 17M_I$, with a standard

deviation of $\sigma(M_T) \simeq 2M_I$ - recall, this is almost independent of δ_t. Percentiles for M_n, and so for M_T, can be obtained through conditional probability calculation: for example, with $\delta_t = 0.5$ and under the latter setting (i.e. $n = 25 \times 85 \times 50\%/(5+1) \simeq 177$), the 80% and 90% percentiles are $m^8_{177} = 122$ and $m^9_{177} = 151$. Once again, percentiles present small variations in function of δ_t.
Mean, standard deviation and percentiles of M_T for the four settings considered of main interest are reported in Table 3. In the light of these further results, even allocation 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 stored) and $r = 25\%$ of resources are allocated to phase II, the average of M_T is $11.1M_I$ and M_T does not overcome $12.8M_I$ with 80% probability, where $OSP(25\%) = 72.5\%$.

Table 3	Standardized measures of total expenses							
k	r	$E(M_T)$	$\sigma(M_T)$	$m^{.8}_n$	$m^{.9}_n$	$m^{.8}_T$	$m^{.9}_T$	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	$r = 25\%$	$11.1M_I$	$3.1M_I$	$1.6M_I$	$2.2M_I$	$12.8M_I$	$15.2M_I$	72.5%
	$r = 50\%$	$16.9M_I$	$2.0M_I$	$1.4M_I$	$1.8M_I$	$18.2M_I$	$19.6M_I$	76.2%

Table 3. 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 $r = 25\%, 50\%$; $\delta_t = 0.5$ has been adopted.

6 Example

Let's continue the introductory example, where the sample size of each phase II group was $n = 60$. Assume here that $h = 7$ doses are studied in phase II, and that two phase III trials are launched if $d_{60} > \delta_{0L} = 0.15$. Setting $\alpha = 2.5\%$ and $1-\beta = 90\%$, if $\delta_t = 0.4$ then $M_I = 132$; also, $m_{\max} = 940$. Note that if δ_t was 0.4, to obtain an observed value of d_{60} near 0.15, and so a phase III sample size estimate close to 940, is not a low probability event, since 0.15 is approximately the 8.5th percentile of $d_{60} \sim N(0.4, 2/60)$.
Considering $w = 20$ times 132 (i.e. $k = 20$ times the ideal phase III sample size), the maximum of OSP is just 67.8% ($r = 47\%$). Now, assume that w is increased up to

$= 25$, in accordance with Table 1, so that resources for treating a total of $25 \times 132 = 3300$ patients are available for the allocation into the two phases. Then, things go better: with r from 29% to 68% the OSP is higher than 70%. In detail, $\max\{OSP\} = 73.5\%$ with $r = 51\%$, where the SP of phase III (also called Average Power) is 76.4% and the launch probability is 96.2%.
This best $r = 51\%$ gives $n = 211$ (i.e. 1688 patients to be enrolled in phase II) and a maximum phase III sample size, per group, of 403. Actual values of phase III sample size result often lower than 403: the average of M_{211} is 146.85, and its standard deviation is 69.74. Consequently, the average and the standard deviation of the total sample size M_T are 2276 and 279. This corresponds to, about, $17.2M_I$ and $2.1M_I$, respectively, meaning that not all the $w = 25M_I$ resources would be spent.

7 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. In particular, resources to develop the research project are often funded for each phase separately. It is a fact that the failure rate of phase II and phase III clinical trials is quite high.
Here, the assumption is that the whole amount of resources to develop phase II and III trials (in terms of sample size) is stored, and therefore potentially available before starting phase II. We studied the problem of allocating the resources to the two phases - to be precise, resources allocated to phase II are all used, where those used in phase III are at most those left, depending on phase III sample size estimation based on phase II data.
It was assumed that 2 phase III trials are run with a sample size estimated on the basis of phase II data. The overall success probability (OSP) has been evaluated as a tool for planning experiments, in accordance with some recent papers [8, 4, 3], and the variability of the resources actually spent has been accounted for.
We showed that to obtain a sufficiently high OSP (e.g. 75%) when the number of doses evaluated in phase II goes from 3 to 9, the whole amount of resources needed varies (linearly) from 19 to 31 times M_I. This is almost regardless 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%. Even an amount of resources of 25% might give a good OSP and an invitingly small total sample size if allocated to phase II, provided that a sufficient amount of resources is stored to the two phases. If the whole amount of resources available for the two phases is low, the OSP will be low too, even lower than 50%, even if the best allocation of resources is made. 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 may be computed by applying assurance [9] on effect size assumptions.

The observed phase II effect size was adopted to compute phase III sample size: being aware of the variability in effect size estimation, conservative sample size estimation strategies may be adopted, as in [1]. The OSP can, therefore, result in a considerable increase (i.e. about 3% when $OSP \simeq 75\%$ - unpublished result). Allocations near 50% providing the optimal OSP are 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 [1]. To improve the success rate of phase II and phase III trials, the drug development could be looked at in its entirety, and 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 profits, of course in case of successful trials. Allocation of resources should also be evaluated from an economic perspective, as suggested also by Jiang [4]. For this reason, our future works may focus on the relationship among allocations, OSP, efficacy and safety utility functions, costs, revenues, and profits, according to [10, 11].

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 [12] are much more evident, since phase II data are considered only for detecting a certain effect with low power, not for adequately planning phase III.

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 eect 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 Eect 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] 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.

[11] 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

[12] Stallard N. Optimal sample sizes for phase II clinical trials and pilot studies. Statistics in Medicine 2012; 31: 1031-1042

*