Published in final edited form as:

Stat Biopharm Res. 2013 April 4; 5(4): 282–292. doi:10.1080/19466315.2013.783504.

Adaptive Budgets in Clinical Trials *

Martin Posch and Peter Bauer[†]

Abstract

We consider situations where a drug developer gets access to additional financial resources when a promising result has been observed in a pre-planned interim analysis during a clinical trial which should lead to the registration of the drug. First the option that the drug developer completely puts the additional resources into increasing the second stage sample size has been investigated. If investors invest the more the larger the observed interim effect, this may not be a reasonable strategy: Then additional sample sizes are applied when the conditional power is already very large and hardly any impact on the overall power can be expected. Nevertheless, further reducing the type II error rate in promising situations may be of interest for a drug developer. In a second step, sample size was based on a utility function including the reward of registration (which was allowed to depend on the observed effect size at the end of the trial) and sampling costs. Utility as a function of the sample size may have more than one local maximum, one of them at the lowest per group sample size. For small effects an optimal strategy could be to apply the smallest sample size accepted by regulators.

Keywords

Interim analysis; adaptive budget; sample size reassessment; reward; sampling costs; utility; power

1 Background

Since the introduction of the statistical testing paradigm there has been a never ending effort to construct designs with certain optimality properties. Adaptive designs started to open a path to flexibility allowing design modification in ongoing trials without compromising on the Type I error rate of the adaptive statistical test (Bauer, 1989; Bauer and Köhne, 1994; Brannath et al., 2002; Lehmacher and Wassmer, 1999; Cui et al., 1999; Proschan and Hunsberger, 1995; Müller and Schäfer, 2001, 2004). Sample size modification has been one of the adaptations which always has attracted interest of experimenters. In particular since it has been realized that in the z-test scenario sample size can be increased if there are promising results in an interim analysis without violating the type I error rate of the conventional test applied in the end (Posch et al., 2003; Blondiaux and Derobert, 2009; Chen et al., 2004; Dunnigan and King, 2010; Mehta and Pocock, 2010; Glimm, 2012). Note that on the other hand it also has been worked out how large the type I error rate inflation may become if the experimenter adopts a worst case strategy of maximizing the conditional type

*We wish to thank Franz Koenig for valuable comments. This work was supported by the Austrian Science Fund (FWF): P23167.

[†]Correspondence to: Peter Bauer, Section of Medical Statistics, Medical University of Vienna, Spitalgasse 23, A-1090 Wien, Austria, peter.bauer@meduniwien.ac.at.

I error rate in any point of the interim sample space without (Proschan and Hunsberger, 1995) or with (Graf and Bauer, 2011) possible adaptation of the treatment allocation ratios at interim.

Strong arguments have been brought up against the use of adaptive test statistics in general (Tsiatis and Mehta, 2003; Burman and Sonesson, 2006; Jennison and Turnbull, 2003) and also against the specific strategy of increasing sample size in case of promising interim results (Emerson et al., 2011). Here we want to look a little closer into a specific scenario described as Example (ii) in Brannath et al. (2006): “In another example new financial resources (available, e.g., due to a promising interim result on the primary or secondary endpoints) may allow for an increase of the overall or conditional power by enlarging the maximum sample size thereby increasing the chance for a successful trial with more precise effect estimates.” We are aware that Bob O’Neill from the position of a regulatory statistician has good reason to argue that interim results from clinical drug trials should not be disseminated as wide that financially interested people play any role in any adaptations of the trial (which is ruled out in regulatory documents EMA, FDA). However such a situation may happen naturally in the development process of a medicinal product, e.g., when very promising results from Phase II may (and should) have an influence on the planning of the Phase III part. Optimizing drug development programs is attracting more and more interest in the biostatistics community as can be seen from several special sessions of the joint conference of the European Medicines Agency, the ISBS and the German region of the IBS in Berlin, 2011. Our situation somehow mimics the above situation with a single trial and an interim analysis (thus looking at the scenario where a single pivotal trial reaching statistical significance would be sufficient for a positive approval by drug regulatory agencies). To even sharpen the arguments there are small companies dealing with a very small number of products which may have a different chance-risk structure than large companies which may rather rely on some sort of long run optimality criteria. We will ask what can be done in a situation when the total budget of a clinical trial depends on results from the trial itself, for simplicity here being a function of the effect estimate found in an interim analysis. A small company may have a limited budget when initiating a clinical trial and, as a consequence, may tend to initiate an underpowered trial based on too optimistic effect sizes. Instead, it could use an adaptive sample size reestimation strategy, with a rather small first stage for which the funding suffices (see also the discussion in Mehta and Pocock (2011)), and count on additional financial support in case of promising interim results. We assume that by the reasoning of investors about risk the additional financial support to be invested (and hence the total budget becoming available for the drug developer) is a monotonic (increasing) function of the interim effect estimate balanced by some compensation from potential future profits for the investor. For the moment we assume such a function is given but do not look into the complicated problem how it should be chosen to please the different parties having their interest in such a process. For the beginning we distinguish between two extreme strategies:

1. The trial has been planned appropriately, no changes are made, the additional budget is put in other tasks (e.g., intensifying the preparation of later marketing because there has been already one successful pivotal trial). This is in line with conventional strategies. One problem is the condition behind “the trial has been

planned appropriately” or more specifically “has been powered in a realistic way”. To our own regulatory experience, for rare diseases trials are often planned with overly optimistic effect sizes in order to get them on the track with “manageable” sample sizes. The other problem arises from the existing conventions on the type II error rate. Whereas sponsors will be reluctant to sharpen the two-sided 0.05 significance level in the two pivotal studies paradigm, they may be not at all be satisfied by a type II error as large as 0.2 or even of 0.1 even if the study is powered on appropriate effect sizes.

2. The additional budget is put into sample size increase. This strategy (ABIS, Additional Budget Into Sample) would decrease the type II error rate at the effect size used for planning and seems to be plausible at the first glance: If there is an indication of a large beneficial effect I try to reduce the risk of missing it at the end. At the second view this strategy creates another problem: Small effects (possibly irrelevant effects in relation to the existing risk of the therapy) will tend to show up as significant with higher probability. But here one may argue that the larger sample size would allow a more precise estimation which then could facilitate balancing risk and benefit (here not addressing the bias which may be introduced by such a data dependent sample size reassessment strategy).

A general note is due here. Mid-trial sample size reassessment to control a prefixed conditional power (based either on the effect size from the planning phase or the interim effect estimate) would reduce the sample size the more the larger the observed interim effect, putting aside for the moment the influence of such a strategy on type I error rate. Is it reasonable to base sample size reassessment on an interim estimate of the conditional type II error rate which then is fixed to some value? It has been shown that conditional power is a highly variable quantity in particular if calculated too early in a trial. When using the observed effect size for powering this is more serious since the interim estimate is used twice, first plugged in as the effect size and second in the interim test statistics the power is conditioned on (Bauer and Koenig, 2006). E.g., Mehta & Pocock 2010 in their practical proposal increase the sample size only in a zone of moderate observed effect sizes and do not reduce the sample size if the effect estimates are small or large. Within the middle zone, however, the sample size as a function of the effect size is monotonically decreasing such that the conditional power is controlled. Being far away from having experience with sophisticated financing arguments in drug development we are open minded enough to concede that without relying on ideology a quick and convincing answer how to proceed with additional resources accessible after good interim outcomes is anything but easy. Before going on, the readers may check for themselves how (and under which side conditions) he may decide in such a situation.

2 The ABIS Strategies

Consider a setting where a single two armed controlled confirmative trial showing a significant treatment effect at one-sided level $\alpha = 0.025$ is required to bring a drug to the market. We assume that a z-test for the comparison of means of normally distributed endpoints is performed. For simplicity we assume that the variance is known and set w.l.o.g.,

$\sigma = 1$. It is aimed to achieve a power of 80% at an effect size of δ_0 , leading to a pre-planned per group sample size of n . We assume that the cost of the trial is linear in the sample size, with a cost c per patient, such that the overall cost is $B = 2cn$. Without limiting generality, we set $c = 1$ (expressing the budget B in units of c). We assume the company has only a limited budget $B_1 < B$ and therefore plans the first stage of the trial with the available budget leading to a first stage per group sample size of $n_1 = B_1/2$. The company hopes that the observed efficacy and safety profile at interim will keep things going either by new internal or external resources. In the interim analysis, the effect size $\hat{\delta}_1$ is observed and a second stage budget $B(\hat{\delta}_1)$ becomes available, which will depend monotonically increasing on the interim effect size estimate. The second stage budget determines the per group sample size in the second stage, $\tilde{n}_2 = B_2(\hat{\delta}_1) / 2$.

Since design adaptations are pending, instead of the z-test for the pooled sample the weighted inverse normal method Lehmacher and Wassmer (1999) is planned to be applied to control the Type I error rate. Let $z_1 = \sqrt{n_1/2} \bar{x}_1, z_2 = \sqrt{\tilde{n}_2/2} \bar{x}_2$ denote stage-wise z-statistics, where $\bar{x}_1, \bar{x}_2$ denote the sample means calculated for each stage separately. Then in the final analysis, the null hypothesis of no treatment effect is rejected if

$$\sqrt{n_1/n} z_1 + \sqrt{1-n_1/n} z_2 > z_{1-\alpha}$$

where $z_{1-\alpha}$ denotes the $1 - \alpha$ quantile of the standard normal distribution.

Figure 1a shows the resulting sample size reassessment strategy and predictive power (the conditional power given the first stage data under the assumption that the observed interim effect is the true effect) for the specific example of a trial powered at 80% at the effect size $\delta_0 = 0.25$. This effect size would imply a sample size of $n = 252$ per group. However, the budget allows only for a trial half this size, such that the first stage per group sample size is $n_1 = 126$. Assume the budget for the second stage is given by

$$B_2(\hat{\delta}_1) = 2 \min \left[\left(n - n_1 \right) \frac{\max(0, \hat{\delta}_1)}{\delta_0}, 2 \left(n - n_1 \right) \right].$$

Thus, if the interim effect estimate $\hat{\delta}_1$ is equal to the value δ_0 used in the planning of the study, the pre-planned sample size will be used. Otherwise the second stage sample size is a truncated linear function such that the maximum second stage sample size does not exceed twice the pre-planned second stage sample size. Note that by this assumption the study is stopped for futility when the observed interim effect is negative. Interestingly, using this strategy, the overall power is (given $\delta = \delta_0$, as originally planned) only 74% and thus below the target power of 80%. This can be explained by the reduction of the sample size in case of small observed effects at interim (and by stopping for futility) which is not compensated by the increase of the sample size for large observed effects (where the chance of a rejection in the final analysis is large anyway). The expected per group sample size is 252, the same as the sample size of the fixed sample test with 80% power. Figure 1b compares the power and

expected second stage sample size with a fixed sample test that has a futility stopping rule if the interim effect estimate is negative. For all relevant effect sizes the power of the ABIS test is somewhat lower than for the fixed sample test. The expected second stage sample size of the fixed sample test with futility stopping is lower if $\delta > \delta_0$ and somewhat larger for $\delta < \delta_0$. This is a consequence of the adaptive budget which leads to sample sizes lower than the pre-planned if $\hat{\delta}_1 < \delta_0$ but larger sample sizes otherwise.

In Figure 2 an alternative ABIS rule is considered which uses the pre-planned sample size if the effect size is larger than zero but lower than the pre-planned δ_0 and a linear increase in sample size otherwise. Again as in the previous example the trial is stopped for futility if a negative treatment effect is observed at interim and the adapted second stage sample is constrained not to exceed twice the pre-planned second stage sample size. Here the power of the fixed sample test (with futility stopping) is practically identical to the ABIS strategy, while the expected sample size of the ABIS strategy is considerably larger for $\delta > \delta_0$.

Note that if $z_1 \geq z_{1-\alpha\sqrt{n_1/n}}$ (which in the example holds if for the interim effect size $\hat{\delta}_1 \geq 0.17$) after a sample size increase (compared to the pre-planned sample size n) the standard z-test statistics and critical value can be used instead of the inverse normal test. Similarly if $z_1 \leq z_{1-\alpha\sqrt{n_1/n}}$ (which in the example holds if for the interim effect size $\hat{\delta}_1 \leq 0.17$) after a sample size decrease (compared to the pre-planned sample size n) the standard z-test statistics and critical value can be used instead of the inverse normal test. The use of the standard test statistics in these cases will lead to a strictly more conservative test procedure. This follows, because the conditional rejection probability of the standard test is strictly smaller than that of the inverse normal combination test in these cases (see, (Posch et al., 2003; Blondiaux and Derobert, 2009; Chen et al., 2004; Dunnigan and King, 2010; Mehta and Pocock, 2010; Glimm, 2012)).

3 Optimizing expected utility

In this section we consider a utility based sample size adaptation rule. Instead of basing the sample size on the available budget or a (ultimately arbitrarily fixed) conditional power we assume that the investor aims to maximize expected gain, given the data in the interim analysis. The *expected* gain is important for investors, as they have typically a portfolio of investments and gains are averaged.

Let g denote the reward that can be obtained with the drug. We assume the reward is 0 if the null hypothesis cannot be rejected and monotonously increasing in the observed point estimate $\hat{\delta} = (n_1\hat{\delta}_1 + \tilde{n}_2\hat{\delta}_2) / (n_1 + \tilde{n}_2)$ (pooled over stages), where $\hat{\delta}_2$ denotes the observed effect size estimated from the second stage data. This reflects the common practice to assess the clinical relevance of the effect size of a drug (once the null hypothesis has been rejected) based on the point estimate of the treatment effect. Furthermore, the reward may depend on the adapted second stage sample size n_2 : larger sample sizes may imply longer trial durations leading to less time to exclusively market the drug. Thus, if the hypothesis test is performed using the inverse normal method, we can express the reward as function

$$g\left(\hat{\delta}_1, \hat{\delta}_2, n_2\right) = m\left(\hat{\delta}, n_2\right) \mathbf{1}_{\left\{\sqrt{\frac{n_1}{n}} \hat{\delta}_1 \sqrt{n_1} + \sqrt{1 - \frac{n_1}{n}} \hat{\delta}_2 \sqrt{n_2} \geq z_{1-\alpha}\right\}}, \quad (1)$$

where $m(\hat{\delta}, n_2)$ denotes the reward given the observed treatment effect is $\hat{\delta}$ and $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. Note that g also depends on n_1 and the pre-planned sample size n (which we omitted for brevity).

Assuming that the clinical trial costs are linear in the sample size, this leads to the utility function

$$g\left(\hat{\delta}_1, \hat{\delta}_2, n_2\right) - 2cn_2. \quad (2)$$

As above, we can normalize the utility function setting $c = 1$, such that utility is expressed in units of per patient clinical trial costs. Even though this utility is unknown in the interim analysis (as it depends on the final outcome of the trial), we can compute its conditional expectation given the interim treatment effect $\hat{\delta}_1$ and assuming an underlying true treatment effect δ . The conditional expected utility is given by

$$u\left(n_2, \hat{\delta}_1, \delta\right) = \int_{-\infty}^{\infty} g\left(\hat{\delta}_1, \delta_2, n_2\right) f_{\delta}\left(\delta_2\right) d\delta_2 - 2n_2, \quad (3)$$

where f_{δ} denotes the density function of the second stage treatment effect estimate. If we do not want to specify a specific δ for the computation of the conditional utility, we can take a Bayesian approach and assume a prior distribution π_0 for the treatment effect which, at the time of the interim analysis, is updated to a posterior π_1 . The averaged conditional expected utility is then given by

$$U\left(n_2, \hat{\delta}_1\right) = \int_{-\infty}^{\infty} u\left(n_2, \hat{\delta}_1, \delta\right) \pi_1(\delta) d\delta. \quad (4)$$

3.1 Optimal Single Stage Tests

To obtain a better understanding of the utility based approach we first consider the corresponding utility function in a single stage design. This is a special case of the two stage test, setting $n_1 = 0$ and $n_2 = n$. Then the inverse normal test reduces to the standard z-test and the formulas above do not depend on $\hat{\delta}_1$ anymore. Furthermore, π_1 is equal to the prior distribution π_0 . As a simple example we assume that the reward given rejection of the null hypothesis is constant such that $m(\hat{\delta}, n) = R$. In Figure 3 the expected utility (3) as function of n is depicted for several assumed effect sizes δ and values of R .

In a second step we tried to base sample size on the optimization of a utility function which includes the reward of registration and the sampling costs. We allowed the reward to depend on the effect size observed in the final analysis (at registration). We replaced the observed effect size at the end (which is unknown in the interim analysis) by its conditional distribution, given the data at interim. To clarify the concept we first calculated optimal sample sizes for single stage designs maximizing expected utility under a specific true effect size δ . The utility (as a function of the sample size n) of a single stage design and constant registration reward has a unimodal form (Figures 3a, b). However, for small true effect sizes the optimum of the sample size is at zero. Then the probability of rejection is close to the significance level such that the expected reward is about αR . As there are no costs for patients, this results in a positive utility.

Clearly, for decreasing rewards the region of positive values of n where it is not utile to run a trial is getting larger (Figure 3c). After reaching a maximum the optimal sample size decreases again for very large δ . As expected, the power at the optimal sample size approaches 1 as δ increases and is also increasing in R .

The shape of the utility (as a function of n) changes if the reward depends on the observed effect estimate at registration (for simplicity we used a linear dependency). Assuming that the reward depends on the observed mean effect such that $m(\hat{\delta}, n_2) = 4\hat{\delta}R$ (we add the factor 4 such that for an observed effect size of 0.25 the reward is equal as in the setting above), the utility function and optimal sample size and power are given in Figure 4. Now the curve may have more than one local maximum, one of them at the lowest per group sample size $n = 1$ (Figures 4a, b). This local optimum is a global maximum when the effect size is very small. This can be explained due to the very small sample sizes and the related high variability of the effect estimates: although rejection occurs with a low probability, the effect estimates needs to be very large (and highly biased) to yield statistical significance provoking eventually high rewards. Additionally, the low sample size implies low costs resulting in a higher utility then the reward that can be obtained with an appropriately powered trial. The latter will be very costly in terms of sample size and will likely end up with a more precise (and thus lower) treatment effect estimate. Especially if there is no effect at all, hoping for the false positive with a minimal sample size may be a utile strategy. Since this is only a theoretical solution arising from a simple model in practice for very small effects δ an optimal strategy could be to apply the smallest sample size accepted by regulators. This result is of some relevance in the area of rare diseases or for subpopulations where not always appropriately powered studies are required. Note that for a given reward the optimal sample size (and thus the power at this sample size) is not continuous in δ and jumps to 1 below at certain value of δ (Figures 4c,d). The discontinuity occurs at the δ value where the local maximum at $n = 1$ becomes the global maximum.

The above optimizations were performed for specific alternatives δ . Figure 5 shows the respective calculations when the utility function is averaged over a normal prior π_1 for the treatment effect. For the numeric example we chose a normal prior with variance 0.25^2 . Not surprisingly, the specific behavior of the utility function is considerably smoothed when averaging over a prior on δ (Figure 5a shows the utility functions for a normal prior around

the effect size the study is powered for in the planning phase, 5b demonstrates the impact of the prior mean on the utility function). In general the utility is maximized at lower sample sizes because we also average over scenarios with no or large treatment effects where the optimal sample sizes for the corresponding δ values are small (Figure 5c). The optimized sample size and the Bayesian predictive power are also non-continuous now as a function of the mean of the prior distribution, however, with smaller steps (Figures 5c,d). (As the variance of the prior distribution decreases we again end up with the optimization for a single parameter value).

3.2 Optimal second stage sample size in the two stage tests

Consider now the setting as in Section 2, where a trial is planned to achieve a power of 80% for the standardized effect size $\delta_0 = 0.25$ but the budget allows only to perform a trial half of this size. We assume that at the interim analysis new investors will provide additional resources in order to optimize the conditional expected gain (4).

The difference to the single stage case is that the interim data at the one hand influence the conditional significance level to be applied for the second stage test statistics when formally applying the inverse normal combination test in the final analysis. At the other hand the interim data together with the initial prior determine the posterior distribution at interim. Hence the conditional expected utility (conditioning on the interim outcome) is the one of a single stage test at the conditional significance level based on the posterior distribution at interim. In the considered example, if the reward is assumed to be a (linearly) increasing function of the effect size observed at the end, the optimal second stage sample size is a unimodal function of the interim effect size (being largest at intermediate effect sizes, see Figures 6a and b) and it is monotonically increasing with the means of the initial priors considered (Figures 6c and d) (assuming as above that $m(\hat{\delta}, n_2) = 4\hat{\delta}R$ and a prior with variance $\sigma_{\pi_0}^2 = 0.25^2$ and $R = 2000$). Clearly the influence of this prior is rather low if its variance is large enough to be dominated by the interim data.

4 Discussion

We looked at the situation that a drug developer gets access to additional financial resources when a promising interim result has been observed during the drug development process. To streamline the arguments we focused on the rather unconventional scenario that for achieving registration a single pivotal trial is performed with a planned (adaptive) interim analysis: The (additional) budget becoming available in the interim analysis is depending on the interim effect estimate. Such a situation to our own regulatory experience may occur in the area of rare diseases. The additional budget may come from investors other than the rather small company developing the drug - the future of the company fundamentally depending on the outcome of the registration decision. The first adaptive strategy (ABIS) investigated was to put the additional budget completely into the second stage sample following the interim analysis. This could be motivated by the developers wish by no means to miss an effective drug, particularly when a large treatment effect has been observed at the first stage. The findings for this type of strategy overall are on first sight discouraging. If we

assume that investors will try to keep their own risk low and rather behave in a way to put money into the project only if, e.g., the effect estimate is larger than the effect size the study has been powered for in the planning phase then the ABIS-strategy will hardly increase the power as compared to a conventional non-adaptive design. The simple reason is that for such promising interim effects the conditional power given the interim data is already very large so that overall little can be achieved by piling up additional sample size in very promising cases only. The result is not surprising at all, because a good developer strategy to achieve a certain overall power in an adaptive design with low average sample sizes would tend to increase sample size rather for moderate interim effect estimates (Posch et al., 2003; Mehta and Pocock, 2010). Still, for a small company it may be rational to invest additional resources, as type II error rates being applied conventionally may be too large if the success of the trial decides on the future of the company. However, reducing the type II error rate further, conditional on a promising interim outcome, may not be a reasonable strategy (depending on what has to be given back for the additional budget). Instead, big pharmaceutical companies might try to maximize the probability of being successful (=registration of one or more drugs) over its whole portfolio and not for a single specific trial/drug only. Indeed it has been shown that company size has an impact on the final outcome of drug development (Regnstrom et al. (2010)).

In a second step we tried to base sample size on the optimization of a utility function which includes the expected reward of registration and the sampling costs. We allowed the reward to depend on the effect size observed in the final analysis (at registration). To clarify the concept we first calculated optimal sample sizes for single stage designs maximizing expected utility under a specific true effect size δ . If the reward (as a function of n) depends on the observed effect estimate at registration the utility function may have more than one local maximum, one of them at the lowest per group sample size $n = 1$ where the effect estimate is highly variable resulting in a highly biased effect estimate, conditional on rejection. The extension to the two-stage case gives similar results for the optimization of the second stage sample size. Note that we did not include the risk associated with the consequences of unblinding interim data (a main concern of regulators) into our considerations.

Overall the arguments applied reassure existing results from the literature. It may not be reasonable to put additional financial resources in a running development process by increasing the pre-planned sample sizes when interim results of an appropriately powered process are very promising. Using a simple model for a utility based choice of the sample size after the interim analysis resulted in the known strategy that for very small and very large interim effects rather small second stage sample sizes should be applied. Increasing sample size (as compared to a pre-planned final fixed sample test) is a utile strategy only for intermediate effect sizes which leads to the unimodal function of the optimal second stage sample size. We agree that there are several issues where our simple calculation could be extended: The model applied is very simple including the distributions of the outcome variable and the definition of the utility function. We allowed the reward to depend on the final effect estimate (at registration) because this is what is commonly used when assessing risk and benefit by regulators. However we did not include the variability of the estimate which seems to be the main reason for the bimodality of the utility function for the single

stage design. Another assumption would be that the reward is depending on the true effect size (Berry et al., 2002). We also did not include a certain amount of baseline costs of a clinical trial not depending on sample size and costs for losing time when increasing sample size. Further extensions would be to consider the situation of two pivotal trials and of a portfolio of several drugs. However, we believe that the basic tendencies for the two-stage design will not substantially change when making further steps to more complicated models or when using tests other than the inverse normal combination test. One fundamental issue is that we only looked at the situation that the additional budget becomes available in an unplanned manner, i.e., it has not been a feature to be considered already in the planning phase of the trial. Although such a situation may arise in practice a further research topic would be to investigate the scenario where an outcome dependent financing of the second stage is pre-planned from the very beginning, i.e., including the first stage design into the optimization process (if and under which conditions needed at all).

References

- Bauer P. Multistage Testing with Adaptive Designs (with Discussion). *Biometrie und Informatik in Medizin und Biologie*. 1989; 20:130–148.
- Bauer P, Koenig F. The reassessment of trial perspectives from interim data – a critical view. *Statistics in Medicine*. 2006; 25:23–36. [PubMed: 16220517]
- Bauer P, Köhne K. Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics*. 1994; 50:1029–1041. [PubMed: 7786985]
- Berry DA, Müller P, Grieve AP, Smith MK, Parke T, Blazek R, Mitchard N, Krams M. Adaptive Bayesian designs for dose-ranging trials. In: Carlin B, Carriquiry A, Gatsonis C, Gelman A, Kass R, Verdinelli I, West M, editors. *Case Studies in Bayesian Statistics V*. Springer; New York: 2002. 99–181.
- Blondiaux E, Derobert E. A New Method for a One-Shot Unblinded Sample Size Reassessment in Two-Group Trials: How & When?. 2009. <http://mat.izt.uam.mx/profs/anovikov/data/TWSM2009/contributed>
- Brannath W, Bauer P, Posch M. On the efficiency of adaptive designs for flexible interim decisions in clinical trials. *Journal of Statistical Planning and Inference*. 2006; 136:1956–1961.
- Brannath W, Posch M, Bauer P. Recursive Combination Tests. *Journal of the American Statistical Association*. 2002; 97:236–244.
- Burman C-F, Sonesson C. Are flexible designs sound? *Biometrics*. 2006; 62:664–683. [PubMed: 16984302]
- Chen YHJ, DeMets D, Lan GKK. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine*. 2004; 23:1023–1038. [PubMed: 15057876]
- Cui L, Hung HMJ, Wang S. Modification of Sample Size in Group Sequential Clinical Trials. *Biometrics*. 1999; 55:853–857. [PubMed: 11315017]
- Dunnigan K, King D. Increasing the sample size at interim for a two-sample experiment without Type I error inflation. *Pharmaceutical Statistics*. 2010; 9:280–287. [PubMed: 19764040]
- Emerson SS, Levin GP, Emerson SC. Comments on 'Adaptive increase in sample size when interim results are promising: a practical guide with examples'. *Stat Med*. 2011; 30:3285–301. [PubMed: 22105691]
- European Medicines Agency. International Society for Biopharmaceutical Statistics. German Region of the International Biometric Society. *The Second International Symposium on Biopharmaceutical Statistics*. 2011.
- Glimm E. Comments on 'Adaptive increase in sample size when interim results are promising: a practical guide with examples' by C. R. Mehta and S. J. Pocock. *Stat Med*. 2012; 31:98–9. [PubMed: 22213002]

- Graf AC, Bauer P. Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look. *Statistics in Medicine*. 2011; 30:1637–1647. [PubMed: 21495058]
- Jennison C, Turnbull B. Mid-Course Sample Size Modification in Clinical Trials Based on the Observed Treatment Effect. *Statistics in Medicine*. 2003; 22:971–993. [PubMed: 12627413]
- Lehmacher W, Wassmer G. Adaptive Sample Size Calculations in Group Sequential Trials. *Biometrics*. 1999; 55:1286–1290. [PubMed: 11315085]
- Mehta C, Pocock S. Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine*. 2010; 30:3267–3284. [PubMed: 22105690]
- Mehta C, Pocock S. Authors' response to "Comment on adaptive increase in sample size when interim results are promising". *Statistics in Medicine*. 2011; 30:3302–3303.
- Müller H-H, Schäfer H. Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical Group Sequential Approaches. *Biometrics*. 2001; 57:886–891. [PubMed: 11550941]
- Müller HH, Schäfer H. A General Statistical Principle for Changing a Design Any Time During the Course of a Trial. *Statistics in Medicine*. 2004; 23:2497–2508. [PubMed: 15287080]
- Posch M, Bauer P, Brannath W. Issues in Designing Flexible Trials. *Statistics in Medicine*. 2003; 23:953–969.
- Proschan MA, Hunsberger SA. Designed Extension of Studies Based on Conditional Power. *Biometrics*. 1995; 51:1315–1324. [PubMed: 8589224]
- Regnstrom J, Koenig F, Aronsson B, Reimer T, Svendsen K, Tsigkos S, Flamion B, Eichler H-G, Vamvakas S. Factors associated with success of market authorisation applications for pharmaceutical drugs submitted to the European Medicines Agency. *Eur. J. Clin. Pharmacol.* 2010; 66:39–48. [PubMed: 19936724]
- Tsiatis AA, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*. 2003; 90:367–378.

Figure 1.

ABIS strategy. a) The per group second stage sample size and predictive power (black solid and dashed lines) of the ABIS design and the predictive power of the fixed sample test including the futility rule (blue dashed line) as function of $\hat{\delta}_1$. b) The average second stage per group sample size and power of the ABIS design (black solid and dashed lines), and the expected second stage sample size and power of the fixed sample design with futility stopping (blue solid and dashed lines). The grid lines indicate the pre-planned second stage sample size n_2 , the preplanned power of 80% and the effect size δ_0 used in the planning phase.

Figure 2.
Alternative ABIS strategy. (Legend see Figure 1.)

Figure 3.

Single stage design assuming that $m(\hat{\delta}, n_2) = R$ is constant. The expected utility (3) for the single stage design as function of n , a) assuming the true effect size is $\delta = 0.25$ and rewards $R = 3000$ (upper curve), 2000 (middle curve), 1000 (lower curve) and b) assuming reward $R = 2000$ and effect sizes (from top to bottom) of $\delta = 0.5, .25, .15, .1$. c) and d) the optimal n and resulting power as function of the true effect size δ and for $R=3000, 2000, 1000$ (from top to bottom).

Figure 4.

Single stage design assuming that $m(\hat{\delta}, n_2) = 4\hat{\delta}R$. Legend see Figure 3 but note the different scales.

Figure 5.

Single stage design assuming that $m(\hat{\delta}, n_2) = 4\hat{\delta}R$ and a normal prior with variance 0.25^2 . The expected utility (4) for the single stage design as function of n , a) assuming a prior mean $\mu_{\pi_0} = 0.25$ and rewards $R = 3000$ (upper curve), 2000 (middle curve), 1000 (lower curve) and b) assuming reward $R = 2000$ and prior means (from top to bottom) of $\mu_{\pi_0} = 0.5, .25, .15, .1, 0$. c) and d) the optimal n and resulting power as function of the mean of the prior μ_{π_0} and for $R=3000, 2000, 1000$ (from top to bottom).

Figure 6.

Two stage design assuming that $m(\hat{\delta}, n_2) = 4\hat{\delta}R$ and a normal prior with variance $\sigma_{\pi_0}^2 = 0.25^2$ and reward $R = 2000$. The expected utility (4) for the single stage design as function of n_2 , a) assuming prior means $\mu_{\pi_0} = 0.25$ and b) $\mu_{\pi_0} = 0$ for (from top to bottom) interim estimates $\hat{\delta}_1 = 0.5, 0.25, 0.15, 0.1, 0$. c) and d) show the optimal n and resulting power as function of $\hat{\delta}_1$ for the priors $\mu_{\pi_0} = 0.5, .25, .15, .1, 0$ (from top to bottom).