Featured Article

Statistics in Medicine

Received 2 September 2014, Accepted 19 February 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6472

Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls

Peter Bauer,[a] Frank Bretz,[b,c] Vladimir Dragalin,[d] Franz König[a] and Gernot Wassmer[e,f*†]

'Multistage testing with adaptive designs' was the title of an article by Peter Bauer that appeared 1989 in the German journal *Biometrie und Informatik in Medizin und Biologie*. The journal does not exist anymore but the methodology found widespread interest in the scientific community over the past 25 years. The use of such multistage adaptive designs raised many controversial discussions from the beginning on, especially after the publication by Bauer and Köhne 1994 in *Biometrics*: Broad enthusiasm about potential applications of such designs faced critical positions regarding their statistical efficiency. Despite, or possibly because of, this controversy, the methodology and its areas of applications grew steadily over the years, with significant contributions from statisticians working in academia, industry and agencies around the world. In the meantime, such type of adaptive designs have become the subject of two major regulatory guidance documents in the US and Europe and the field is still evolving. Developments are particularly noteworthy in the most important applications of adaptive designs, including sample size reassessment, treatment selection procedures, and population enrichment designs. In this article, we summarize the developments over the past 25 years from different perspectives. We provide a historical overview of the early days, review the key methodological concepts and summarize regulatory and industry perspectives on such designs. Then, we illustrate the application of adaptive designs with three case studies, including unblinded sample size reassessment, adaptive treatment selection, and adaptive endpoint selection. We also discuss the availability of software for evaluating and performing such designs. We conclude with a critical review of how expectations from the beginning were fulfilled, and – if not – discuss potential reasons why this did not happen. © 2015 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: adaptive design; clinical trials; group sequential designs

1. Introduction

With the publication of [1] 25 years ago, the first attempt was made to establish confirmatory adaptive methodologies allowing for flexible mid-trial design modifications in ongoing trials using any available internal (unblinded) and external data without compromising on the type I error rate. The methodology became more widely known with the publication of [2] 5 years later. Although the methodology was intensively discussed from early on, often also very controversially, it took a few more years until it reached broad interest across the clinical trial community [3]. The development of adaptive design methodology can be characterized by several waves of research: In the early days, the major focus was on sample size reassessment, followed from 1999 on by treatment selection and multiple testing [4].

[a] *Section of Medical Statistics, Medical University of Vienna, Spitalgasse 23, 1090 Wien, Austria*
[b] *Novartis Pharma AG, Lichtstrasse 35, 4002 Basel, Switzerland*
[c] *Shanghai University of Finance and Economics, China*
[d] *Johnson and Johnson, 1400 McKean Rd, Spring House, PA, 19477, U.S.A.*
[e] *Aptiv Solutions, an ICON plc company. Robert-Perthel-Str. 77a, 50739, Köln, Germany*
[f] *Institute for Medical Statistics, Informatics and Epidemiology, University of Cologne 50924 Köln, Germany*
[*] *Correspondence to: Gernot Wassmer, Institute for Medical Statistics, Informatics and Epidemiology, University of Cologne 50924 Köln, Germany*
[†] *E-mail: gernot.wassmer@uni-koeln.de*

An important stimulus for further methodological research and its applications [5] was driven by the release of regulatory guidance documents in Europe [6] and the Unites States [7]. Since their publications there seems to be a clear tendency to focus on simpler adaptive designs with a single interim analysis because of the logistic complications of adaptive interim analyses and the intention to maintain scientific integrity and persuasiveness of a confirmatory clinical trial. Following the general trend toward personalized medicine, adaptive subgroup selection designs have attracted considerable interest more recently.

The use of methods without knowledge of the treatment assignments at the adaptive interim analysis (so-called blinded interim analysis) has been developed in parallel as another branch of flexibility [8, 9] which interestingly started more or less at the same time as unblinded methods. Such methods are attractive to many trialists as they may use conventional test statistics, decision boundaries, and confidence intervals at the final analysis [10]. These methods still attract further methodological research, for example [11–13].

In the beginning, the already well-established group sequential community was reluctant to accept connections between adaptive and group sequential designs. There was a sharp controversial discussion on the scientific value of applying adaptive designs in contrast to group sequential designs [14–18] mainly because of the violation of the sufficiency principle [19]. Moreover, this critical view stated that classical group sequential designs are flexible and adaptive enough, in the sense that the sample size is changed in a data-driven way through the introduction of stopping rules. Interestingly, although the arguments have not changed, some of the authors having been very critical initially are nowadays accepting adaptive designs. Today, it is commonly accepted that group sequential trials can be considered as a special case of the more general adaptive designs. That is, any group sequential designs can be made more flexible by using the inverse normal combination function [20, 21] or the conditional error rate principle [22–24].

Many other developments to enhance the flexibility and efficiency of clinical trial designs have attracted considerable attention over the last decades. In this review paper, we will focus on confirmatory adaptive designs in the sense of [1, 2]. Many other important developments exist, such as blinded sample size reassessment [10], Bayesian adaptive designs [25], response-adaptive randomization [26], and type I error rate control by exact calculation [27–30] or simulations [28, 31, 32] of critical boundaries for prespecified set of adaptation rules. However, whether using simulations is sufficient 'to prove' type I error rate control without any doubts remains controversial [33].

In the following section, we provide an overview over the historical developments of adaptive design methodology. In Section 3, we sketch the underlying main statistical principles followed by recent developments in multi-arm designs, adaptive enrichment designs, and testing multiple endpoints. Specific problems in estimation, in adaptive designs with time to event endpoints, and some other challenges are discussed as well. In Section 4, we discuss the perspectives from the pharmaceutical industry and regulatory agencies on confirmatory adaptive designs. In Section 5, we describe three selected case studies on sample size reassessment, treatment selection, and adaptations in pediatric trials, respectively. An overview of available non-commercial and commercial software is given in Section 6 followed by a critical summary on how expectations from the beginning were fulfilled and which of them were not.

2. A brief history of the early days

In the 1980s, the use of post-hoc meta-analysis of more than one experiment attracted increasing attention in medical statistics [34]. The obvious motivation behind it was that decisions on experimental therapies in general had been (and still are) based on series of medical experiments. Such meta-analyses almost exclusively were based on a cumulating series of trials with an unspecified rule when to stop the series. Bauer [35], in contrast, proposed something like a prospective meta-analysis in the context of sequential testing. For example, one of the pre-assigned test strategies was to plan for a maximum of m trials and to stop early as soon as k trials have produced a positive test at an appropriately chosen level (e.g., two significant out of three planned trials). Another test proposed in this paper was a sequential version of Fisher's combination test of p-values of the individual trials. Still the planning of the series of trials was intended to be done in a conventional prospective way. In the first paper on what is nowadays called confirmatory (frequentist) adaptive designs, Bauer [1] wrote that 'one of the drawbacks of the formulation of the 'sequential meta-analysis' is that the series of trials and the corresponding individual null hypothesis have to be fixed in advance. ... If learning from experience is allowed, it might not be reasonable to restrict experimentation to the estimation and test of a preassigned common parameter in all the consecutive studies'. This first paper considered the different stages as different experiments

and claimed that the combination test principle allows for a selection of the null hypothesis to be tested at the forthcoming stages from a finite number of candidates based on the information collected so far. By some simple arguments, it was shown that conditionally on any possible sequence of hypotheses (or designs) selected the distribution of the individual stagewise p-values under the null hypotheses still will be uniform (or stochastically larger than the uniform distribution). Hence this will also hold true unconditionally. Already in this early attempt, there had been full awareness of the problems that could arise when being faced with the interpretation of the overall test decision in such multi-stage experiments with varying hypotheses: 'It has to be conceded that numerous examples could be constructed, where the formulation of a global null hypothesis would not make sense, because the individual null hypotheses or designs (e.g., with completely different inclusion criteria) do not refer to the same underlying question'. This had also been the major concern of the discussants who – completely uncommon in this journal – had been invited by the co-editor, Joachim Vollmar, to comment on the paper.

A further publication seemed to be necessary to reach a broader readership. Bauer and Köhne [2] began with 'Planning an experiment (e.g., a clinical trial) relies on various information, which is generally not available with the required precision when the study protocol is written' and argued that in a planned interim analysis, 'naturally the investigator would then wish to adapt the study protocol to correct for deficiencies'. Adaptive designs now are more looked at in terms of flexible multi-stage designs and not of meta-analysis. Among the 'numerous strategies for testing the intersection H_0 of two (or k) individual null hypotheses', Fisher's criterion had been used as an example 'because it has good properties and is well known'. Although this is a rather poor justification, stopping boundaries for some two-stage and three-stage procedures had been given. More importantly, many of the problems tackled later had at least been shortly addressed in this paper: how to plan adaptive studies ('All the adaptation performed are documented in a prescheduled amendment of the study.' 'One of the drawbacks of such general methods is the lack of concrete rules for modifying the design appropriately, for example, recalculating the sample size.' 'The study protocol has to describe which types of adaptations are intended'), multiple testing of the individual stagewise null hypotheses (with a mistake later pointed out by G. Hommel and corrected in [2]), how survival type data could be used in such adaptive studies ('The basic concept of stochastic independence between the test statistics from the different stages of a trial need further clarification.'), how to treat two-sided testing, and the possible bias in estimation. It should be mentioned that the rising practice to cover modifications of ongoing trials by unscheduled amendments to the study protocol – mostly without seriously considering the consequences for the validity of the statistical analysis and its interpretation – historically has also been an important motivation to seek methods dealing with design changes in a scientifically valid way. However, considering the new type of methodology and the rather crude argumentation on its inferential basis, it is not surprising that it took a long time until the manuscript was finally accepted, although all formal issues raised had been clarified earlier. Brannath *et al.* [36] later extended the combination test principle to the recursive setting with a flexible number of stages by relying on more stringent mathematical arguments.

Proschan and Hunsberger [22] introduced the concept of the 'conditional error function' to adapt the sample size in an interim analysis. They succeeded in calculating the maximum type I error rate conditional on the observed data at interim that can be produced by always applying a 'worst case' (balanced) sample size reassessment rule at interim. Adjusting for that worst case, you may safely perform any type of balanced sample size reassessment without compromising the overall type I error rate.

Fisher [37] introduced the sequential method of self-designing trials, where prior data in a trial can be used to determine the use and weighting of subsequent observations. This can be carried out until 'all the variance of the test statistics has been used up', and early stopping otherwise is only possible for futility. Hartung [38] later showed that Fisher's combination test with a maximum of K stages and a stopping rule based on non-stochastic curtailment can be looked at as such a procedure. Fisher's combination tests with $K > 2$ stages were also considered in Wassmer [39].

Cui *et al.* [21] and Lehmacher and Wassmer [20] by arguments differing somehow in their generality pointed at the important and most natural way of predefining a combination function: When you are dealing with i.i.d. normal outcome variable with known variance, the final test statistics is just the weighted average of the stagewise z-scores weighted by the square root of the corresponding sample sizes. The 'inverse normal combination function' has the convincing property that when no adaptation has been performed, then the conventional sufficient test statistics as for non-adaptive group sequential designs is applied in the statistical analysis. Lehmacher and Wassmer [20] pointed out at the generality of the approach achieved by transforming stagewise p-values to z-scores.

A further important step forward was made by Müller and Schäfer [23]. They adopted the concept of the conditional error probability as a function of the interim outcome within the framework of group sequential designs imposing the following constraint on possible design adaptations: In any interim analysis, the remainder of the pre-planned group sequential test can be replaced by any other design, which for any interim outcome would later on never produce a larger conditional error rate than the original design. When the conditional error rate can be quantified in the initial design, this opens a large field of flexibility with the advantage that again the conventional test is used if no adaptations are made. Posch and Bauer [40] and Wassmer [41] had already shown that there is a one-to-one correspondence to the combination test principle when choosing the matching combination function (see also Wassmer [42]). Optimal conditional error functions were derived in Brannath and Bauer [43]. Müller and Schäfer [24] made the crucial next step in using the conditional error function principle as a general statistical principle for changing a design any time during the course of a trial. This principle can also be applied in a recursive way, and the conditional error function somehow creates the natural combination function, which is defined at any time by the pre-planned design itself.

As a further development from the early days that had a great impact on a variety of future research, for example, dealing with treatment selection and subgroup selection, is the adoption of the closure principle of multiple testing to adaptive frequentist designs by Bauer and Kieser [4], Kieser *et al.* [44], and Hommel [45]. This allows one to claim that individual doses are better than a control in trials with many-to-one comparisons and treatment selection at interm.

To round up the early days, one of the fundamental questions remaining vague for a long time was the probabilistic foundation of adaptive designs (Liu *et al.* [46] and Liu and Pledger [47]). The crucial difficulty is that adaptations in such designs are random variables themselves. A late answer to the fundamental problems is based on the assumption of 'regression independence' (Brannath *et al.* [48]): Summarizing the author's arguments for adaptive two-stage designs, we need to have a regression model that specifies the distribution of the second stage data, given the interim data and second-stage design. With regard to type I error rate control, this regression model under the null hypothesis must specify a valid conditional distribution for the data to be observed after the adaptive interim analysis and to be used for combined inference. This conditional distribution under the null hypothesis has to be independent from unknown features such as the investigator's choice for interim outcomes different from the one actually observed. The crucial point is that by pre-specifying such a generally valid version of the required conditional probabilities, any data-driven choice and ambiguity of versions is avoided. Consequently, 'artificial constraints that had to be imposed on the investigator only for mathematical tractability of the model are no longer necessary'.

3. Statistical methodology and new developments

In the following, we sketch the basic methodology to perform an adaptive interim analysis based on unblinded data. We restrict arguments to two stages, that is, with a single adaptive interim analysis. The basic principles can be generalized to more than two stages fairly easily. In general, flexibility can be introduced in confirmatory clinical trials by two related approaches: the combination test and the conditional error approach.

3.1. Adaptive designs based on combination functions

Consider a null hypothesis H, which is tested in two stages. Let $p,\ q$ denote stagewise p-values for H, such that p is based only on the first stage and q only on the second stage data. This is the main clue of adaptive combination tests: not to pool the data over stages but to combine the information via stagewise calculated test statistics. Using appropriate invariance principles, that is, that the stagewise p-values for continuous test statistics follow each a uniform distribution under the null hypothesis, the joint distribution of any pre-defined combination function can be derived. A two-stage combination test [1, 2] is defined by a combination function $C(p, q)$, which is monotonically increasing in both arguments, boundaries for early stopping $\alpha_1, \alpha_{0,}$ and a critical value c for the final analysis. The trial is stopped in the interim analysis if $p \leqslant \alpha_1$ (early rejection of the null hypothesis) or $p > \alpha_0$ (early acceptance due to futility). If the trial proceeds to the second stage, the null hypothesis is rejected if $C(p, q) \leqslant c$, where c solves

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{[C(x,y)\leqslant c]} dy\, dx = \alpha \tag{1}$$

Here, the indicator function $\mathbf{1}_{[\cdot]}$ equals 1 if $C(p,q) \leqslant c$ and 0 otherwise. By the definition of c, the combination test is a level α test. This still holds if the design of the second stage (e.g., the sample size and/or test statistic) is based on the interim data. The only requirement is that under H, the distribution of the second stage p-value q conditioned on p is larger than or equal to the uniform distribution [36]. This is, for example, the case when new patients are recruited in the second stage and conservative tests are used at each stage. Examples of combination functions are the product of the p-values $C(p,q) = p\,q$ [1, 2], and the weighted inverse normal combination function [2, 20, 21] defined as $C(p,q) = 1 - \Phi\left[v\,\Phi^{-1}(1-p) + w\,\Phi^{-1}(1-q)\right]$, where v, w denote pre-defined weights such that $v^2 + w^2 = 1$, Φ is the cumulative distribution function of the standard normal distribution and Φ^{-1} its inverse. For the one-sided test of the mean of normally distributed observations with known variance, the inverse normal combination test with stagewise sample sizes n_1, n_2 and weights $v^2 = n_1/(n_1+n_2)$, $w^2 = n_2/(n_1+n_2)$ is equivalent to a classical two-stage group sequential test (the term in squared brackets is simply the standardized mean Z). Note that this equivalence also applies for adaptive designs with more than two stages. Thus, the critical values α_1, α_0, c for the inverse normal method can be computed with standard software for group sequential trials (Section 6).

The process of adaptive designs can be summarized in three important phases: the planning phase, the interim analysis, and the final analysis. In the planning phase, the design of the first stage has to be prespecified including the null hypothesis H, test statistic, sample sizes, and the combination test (with all critical boundaries α_1, α_0, and c). In the adaptive interim analysis, it is first checked whether the trial can be terminated because of crossing an early decision boundary. Otherwise the design of the second stage can be fixed using all information gained. This includes the second stage sample sizes, the null hypothesis H' to be tested at the second stage, and the corresponding test statistic.

The change of the hypothesis H' might lead to a pitfall, if not considered appropriately in the analysis and interpretation. For example, one may pre-define a non-parametric test (e.g., a Wilcoxon test) for the first stage and naively switch to a parametric test (e.g., t-test) for the second stage data, because the data of the first stage support the application of a more powerful parametric test. However, the question is how to interpret a final rejection of the combination test. Due to switching, strictly speaking, two different null hypothesis are tested at the two stages. Hence, one has to be very careful in interpreting a final rejection of the combination test: Strictly only the intersection of $H \cap H'$ has been rejected. Note that no formal inference for the elementary null hypothesis H and H' has been performed. In practice, one has to decide whether such a rejection of the combined global null hypothesis is sufficient for the interpretation of the study results. For further information on how to deal with controlled inference on multiple hypotheses in adaptive designs, we refer to Section 3.3.

3.2. Adaptive designs based on the conditional error function

Another approach to introduce flexibility in design adaptations after an interim analysis is to calculate the probability to reject the null hypothesis H in the originally pre-planned design at level α under the assumption that H is true given the interim data collected so far. This probability $A\left(X_1\right)$ (as a function of the interim data X_1) is called the conditional error function [22]. Now we can replace the remainder of the pre-planned design by any new design which does not produce a conditional error rate for a final rejection of the null hypothesis given the interim data X_1 exceeding $A\left(X_1\right)$. Consequently, if this holds conditionally for any X_1, it holds unconditionally.

$A\left(X_1\right)$ denotes a measurable function from the first stage sample space to the unit interval [0,1] such that $E_H(A) \leqslant \alpha$. In practice, the sample size and test statistics for the second stage are planned based on interim data resulting in a second stage p-value q (based on the data of the second stage, only). Finally, H is rejected if

$$q \leqslant A\left(X_1\right) \tag{2}$$

Note that if $A = 0$ (early acceptance) or $A = 1$ (early rejection), no second stage needs to be performed for the test decision. This procedure controls the type I error rate as long as the conditional distribution of the second stage p-value q, given the first stage data, is under H stochastically larger or equal to the uniform distribution. Proschan and Hunsberger [22] were the first who introduced a conditional error function to address sample size adaptations. Müller and Schäfer [23] showed that in principle, the conditional error function $A\left(X_1\right)$ is defined for any pre-planned design, even if no interim analysis was originally foreseen [24]. Thus, from a pure statistical point of view, even in a fixed sample size design after each single observation, an unplanned adaptive interim analysis based on the preservation of the

conditional error of the originally planned design can be performed without violating the overall type I error rate. This was an important milestone in introducing flexibility into frequentist confirmatory trials.

This feature allows to apply the originally pre-planned test if no adaptations are performed at the adaptive interim analysis. Thus in the interim analysis, one has the option to complete the trial as initially planned or to choose any other test for H (with new observations) at the level of the conditional error function. If adaptations are performed, the null hypothesis H (or $H \cap H'$ in case of a change of the hypothesis) is rejected based on the second stage p-value q whenever (2) is satisfied.

Instead of describing the conditional error function approach in terms of stagewise incremental statistics, an alternative way would be in the style of group sequential designs by presenting simply the cumulative test statistics. One reviewer pointed out that this way it may be 'easier to see that the conditional error approach is a natural extension of the classical group sequential methodology and specializes to it in the absence of any design change.' However, to achieve a strict type I error rate control and to keep the presentation in terms of cumulative test statistics require an adaptation of the critical boundaries in a data dependent way, that is, part of the data observed before the adaptation are 'hidden' in the modified boundaries. This may give the misleading impression that the usual sufficient cumulative test statistics are used for the test decisions equally weighing all observations irrespective of the adaptations. Using such data-dependent boundaries may lead to identical rejection regions as for adaptive combination tests anyway, and the presentation may be a matter of taste.

Essentially, the conditional error function defines itself a combination function so that both approaches are equivalent in principle [40–42]. However, if nuisance parameters are involved, the calculations of the conditional error may become tedious or even impossible [49–51]. This is one of the drawbacks of the elegant conditional error approach. The combination test approach with stagewise test statistics (and related p-values) may be still applicable in a straightforward way. However, it suffers from using a type of overall (pooled) test statistic, which in general one would not use in the final analysis of a conventional design: For example, by using the inverse normal function and stagewise t-tests, estimates of nuisance parameter are calculated separately within each stage for t-statistics using stagewise means. Then the p-values derived from the stagewise t-values are combined into the combination test statistics.

In the remainder of the manuscript, we focus on adaptive combination tests (refer to them as adaptive tests), although the same flexibility can be achieved by the conditional error approach.

3.3. Sample size reassessment

The main motivation in the early days to consider confirmatory adaptive designs as an option was the opportunity of performing unblinded sample size reassessment (SSR) (for example, [20, 21, 52–54]). The main approach of recalculating the sample size at interim is to use conditional power arguments [22, 55, 56]. However, Bauer and König [57] showed that if the interim analysis is performed too early, one may be misguided by the highly variable interim data, especially when the interim estimate of the effect is used twice: Additionally to conditioning on the observed interim effect, it is also used straightaway as the 'true' effect for the calculation of the conditional power. Such a strategy may lead to a highly variable distribution of the conditional power resulting in highly variable sample sizes to achieve a targeted conditional power. Hence, whether and how SSR (timing and method) should be performed has to be considered carefully in advance and latest at interim. Note that SSR is a useful tool, and the method is allowing for it but without recommending or even enforcing its use.

3.4. Adaptive designs and multiplicity adjustments

Confirmatory clinical trials tend to be more complex than simple testing of a single hypothesis. Adaptive designs have been developed to tackle different study objectives at once, that is, testing multiple hypotheses. For the confirmatory clinical trials in drug development, it is required that the probability to reject at least one true hypotheses should be bounded by α irrespective of how many and which null hypotheses are in fact true(for example, [58]). This is referred to as control of the multiple level of the analysis (or of the family wise error rate, FWER, in the strong sense). Such objective may refer to testing primary and secondary endpoints, multiple treatment-control comparisons with interim treatment selection, or subgroup selection and testing. A powerful way to construct tailored multiple testing strategies in adaptive designs is the use of adaptive tests in combination with the closed testing principle [59].

Consider a clinical study where null hypotheses $H_1, \ldots, H_G$ are to be tested. For example, this could denote multiple elementary null hypotheses referring to multiple treatment-control comparisons, multiple

endpoints, or multiple subgroups. The global null hypothesis

$$H_0 = \bigcap_{g=1}^{G} H_g$$

is the intersection of elementary null hypotheses. When performing a closed testing procedure, we first derive the corresponding closed system of hypotheses consisting of all possible intersection hypotheses

$$H_{\mathcal{I}} = \bigcap_{i \in \mathcal{I}} H_i\,, \quad \mathcal{I} \subseteq \{1, \ldots, G\}$$

including all elementary hypotheses H_g, $g = 1, \ldots, G$. Then, for each of the hypotheses in the closed system, we determine a suitable local level-α test. In the simplest case, this is the Bonferroni test, but other tests (e.g., the Simes or Dunnett test for many-to-one comparisons) can also be used. Having observed the data, we reject $H_{\mathcal{I}}$ at FWER α, if all hypotheses $H_{\mathcal{J}}$ with $H_{\mathcal{J}} \subseteq H_{\mathcal{I}}$ are rejected, each at (local) level α.

According to the closed testing principle, an elementary hypothesis H_g can only be rejected if the elementary null hypothesis H_g itself, and all intersection hypotheses containing H_g are rejected, each at level α. This is the key to introduce flexibility. Instead of using standard tests from the non-adaptive setting, a level-α combination test is applied to each intersection hypotheses.

A general problem in applying the closure principle to adaptive design with multiple hypothesis testing is that an adaptive level-α combination test has to be defined for each of the intersection hypotheses. A further problem occurs if at an adaptive interim analysis, one (or more) null hypotheses are dropped from the initial family of hypotheses to be tested. Consider, for example, a clinical trial with three treatment-control comparisons and assume that two treatments are dropped at an interim analysis, and the second stage sample size is reassessed. Thus, only one elementary hypothesis is continued to the second stage and eventually the final analysis. In order to strongly control the FWER at level α according to the closed test, any intersection hypothesis contained in the continued hypothesis has to be rejected in the final analysis at level α. As there are no second-stage data for the null hypotheses dropped at the interim analysis, the question is how to define a second-stage p-value to be used for the combination test of these intersection hypotheses. A valid p-value is the p-value of the single remaining hypothesis containing the intersection hypothesis. As a matter of fact, the second stage p-values for the intersection cannot be fixed in advance but is (and has to be) fixed in the interim analysis (as part of the adaptive second stage design). As already mentioned in Section 3.1, the design of the first stage including the combination function has to be fixed in advance, but not the second stage design.

Such an adaptive approach has been suggested by Bauer and Kieser [4] for testing multiple hypotheses: The multiplicity adjustment for the first stage is an important part of the first stage design and has to be fixed a priori. Because adaptations by the nature of adaptive designs are not pre-specified, the adaptive selection of hypotheses may impact the multiplicity adjustments for the second stage. Therefore, the multiplicity adjustment for the second stage cannot be laid down a-priori in the planning phase of the trial but has to be specified at the interim analysis as part of the second stage design. This opens a further area of flexibility (and complexity), because the first stage data can be used to tailor the multiple testing strategy for the second stage, for example, introducing an order relation for the second stage. It is even possible in an adaptive interim analysis to introduce new hypotheses [45]. However, specifying combination functions, stopping boundaries, stagewise sample sizes, and stagewise multiplicity adjustments for the intersection hypotheses may be a challenging task. How to apply the closed testing principle within adaptive designs was proposed by several authors (for example, [4]). We refer to the tutorial of Bretz *et al.* [60] for an overview on adaptive closed tests and a discussion of potential multiplicity adjusted p-values to be used in the combination test setting. In the following, we describe three applications of adaptive designs with multiple inference.

3.4.1. Many-to-one comparisons. The first application of testing multiple hypotheses were developed for the many-to-one comparisons [4, 45, 61, 62]. This principle can easily be extended to multiple stages and/or to other combination functions, for example, Fisher's combination test and to the conditional error approach. Such multi-arm clinical trial designs with an adaptive interim analysis have also been referred to as adaptive seamless designs [63–67]. Friede and Stallard [68, 69] compared adaptive designs to group sequential designs with treatment selection, whereby just recently, the latter approach has been made flexible using the closed testing and conditional error principle [70], see also Gao *et al.* [71].

Particularly, König *et al.* [72] proposed a procedure that is based on the conditional error of the single-stage Dunnett test. They showed that this procedure uniformly improves the Dunnett test if treatment arms were selected at an interim stage. The test coincides with the classical Dunnett test if no treatment arm selections (or other adaptations) were performed. Application within the closed testing procedure is straightforward. The procedure is different from the inverse normal method, when a Dunnett test is used for testing intersection hypotheses and in most cases more powerful.

3.4.2. Adaptive enrichment designs. In recent years, a general trend toward personalized medicine has stimulated innovative clinical trial designs looking at the treatment effect in the overall study population, but also in targeted subgroups. By enriching a population and focusing on the targeted subgroup only, one runs the risk to restrict an efficacious treatment to a too narrow fraction of a potential benefiting population. However, when using a too broad study population (with an effect in a biomarker positive but no or only a modest effect in the biomarker negative patients), one might fail to prove efficacy because of a dilution of the treatment effect. Therefore, adaptive enrichment designs have been suggested. In the first stage, patients are recruited from the full population. At an adaptive interim analysis, the trial population may be adapted based on the observed treatment effects. The trial continues either in the full population or in a subpopulation only. If the full populations is selected, one may further decide to test both hypotheses (full and subgroup) or the full population only for the remainder of the trial. To control the type I error rate adjusting for the adaptive choice of populations as well as the multiplicity arising from the testing of subgroups, combination tests [73–78] and the conditional error rate principle [79–82] have been proposed. Different decision rules to select the population for the second stage have been considered, ranging from simple rules based on differences of z-statistics [81] to Bayesian decision tools [73, 77]. Generalization to more than one subpopulation was considered in [78].

3.4.3. Multiple endpoints. Clinical trials often address several outcome variables within a single confirmatory experiment, and multiple tests are part of the confirmatory statistical analysis. For example, one may have more than one primary endpoint (Section 5.3) or a mix of primary and secondary endpoints. Here, O'Neill [83] points out, for example, that secondary endpoints shall not be tested before efficacy in the primary endpoint has been shown [84, 85]. When the number of outcome variables is limited, still it might be feasible to set up the full adaptive closed test to get strict error control even for secondary outcome variables. However, due to increase sophistication of multiple testing strategies, the communication of the related multiple testing procedures and its interpretation to clinicians and sometimes even to statistical colleagues gets complex. Hence, graphical methods to visualize the logical structure have been suggested to facilitate planning, execution, and interpretation of such complex multiple tests for conventional fixed size sample designs [86–88]. Recently, adaptive graph-based methods for multiple comparisons based on combination tests [89, 90] or conditional error function [91] have been proposed.

3.5. Further methodological challenges in adaptive designs

As illustrated earlier, the theory of adaptive designs is well understood for immediate responses; there are still problems arising from delayed responses, for example, in survival trials. Schäfer and Müller [92] proposed adaptive survival tests using the independent increments property of logrank test statistics [93–95]. However, Bauer and Posch [96] show that these methods may fail if other data than the interim logrank test statistics are considered when redesigning the study, for example, using progression free survival to predict overall survival times for patients still being alive at interim. Strictly, investigators may not use any data allowing a prediction of survival times of patients alive at interim.

There are several suggestions how to construct valid adaptive test statistics including as much information as possible and allowing interim decision making on all collected unblinded data [74, 97, 98]. The latter two papers elaborate also on subgroup selection. Magirr *et al.* [99] show that these proposals have the common disadvantage that the final test statistic may ignore a substantial subset of the observed survival times. They show that if the goal is to use all the data, a worst-case adjustment of the critical boundaries guarantees type I error rate control for the price of reduced power.

Other methods require assumptions regarding the joint distribution of survival times and short-term secondary endpoints [81, 100–102]. Note that related problems arise in overrunning, for example, patients having been recruited before or during the interim analysis such that their data could not be used for interim decision making [103].

Another important question in adaptive designs is how to derive adequate point estimates and confidence intervals for the treatment effect of interest. This is already challenging in the conventional group sequential setting [104, 105]. The statistical properties of different point and interval estimates in adaptive two-stage designs with sample size reassessment have been investigated [106–108]. Conditional estimates have been suggested in setting of two-arm [109–111] and multi-arm trials [112]. Proposal on confidence intervals developed for group sequential designs have been extended to adaptive designs [113–115]. There are also suggestions for confidence intervals in adaptive designs with treatment selection [62, 116]. Interestingly, in adaptive designs, the bias in multi-arm trials is smaller than in fixed sample designs [117].

Although confirmatory adaptive designs have been developed in context of clinical trial, some basic ideas have been recently applied in other fields, for example, for big-data in the genomic area [118–120].

4. Regulatory and industry perspectives

Confirmatory adaptive designs have attracted substantial interest across industry and major regulatory agencies since the early 2000s. Already in 2006, the European Medicines Agency (EMA) released a draft Reflection Paper titled 'Methodological Issues in Confirmatory Clinical Trials Planned With an Adaptive Design' [6], which was finalized in 2007. According to this guidance, a study design is called adaptive 'if statistical methodology allows the modification of a design element (e.g., sample-size, randomisation ratio, and number of treatment arms) at an interim analysis with full control of the type I error rate.' This document thus provided the first regulatory guidance on adaptive designs and essentially focused on confirmatory trials. It acknowledged potential benefits of adaptive trials, but its emphasis was on caution.

In parallel to the increasing understanding and acceptance of adaptive designs in Europe, the US Food and Drug Administration (FDA) included adaptive designs in its 2006 Critical Path Opportunities List [121] as one potential approach to accelerating the translation of new biomedical discoveries into therapies. A few years later, FDA then released a draft guidance on 'Adaptive Clinical Trials for Drug and Biologics' [7]. This guidance defines an adaptive study as one that 'includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study'. Analyses of accumulating trial data are conducted at planned timepoints and can be performed in a fully blinded manner or in an unblinded manner. Acknowledging that many design modifications are possible, the FDA guidance specifically mentions nine examples of possible adaptations: eligibility criteria, randomization procedure, treatment regimens, sample size (including early termination), concomitant medication, schedule of patient evaluations, and the choice of primary and secondary endpoints, including analytic methods for their evaluation. The draft guidance then recognizes two categories of adaptive designs: those which are generally well-understood and for which established approaches to implementation are available, and others whose properties are less well-understood and for which the community needs to collect more experience, although this distinction seems arbitrary to some extent because there are open questions in both categories. Regulatory statisticians specifically from the FDA continuously contribute to methodological research, for some recent contributions see [122–126], see also [127].

Both the EMA and the FDA guidance documents include helpful definitions for adaptive trials as well as important points to consider before embarking on the design and implementation of an adaptive trial. This includes an absolute requirement for prospectively written standard operating procedures and working processes for implementing adaptive trials as well as the acknowledgment that pharmaceutical companies increasingly more often engage with contract research organizations that have the necessary experience in running such trials. Both the EMA and the FDA guidance documents go far in the endorsement of design modifications based on blinded data review. Although many methods are well established and have been in use for some time, research is still needed to investigate open questions, such as estimation and confidence intervals, type I error rate control of blinded sample size re-assessment in non-inferiority trials, and inefficiently large trials in case of a strong treatment effect

At the same time, industry associations like the Pharmaceutical Research and Manufacturers of America (PhRMA), the European Federation of Pharmaceutical Industries and Associations (EFPIA), and the Japan Pharmaceutical Manufacturers Association (JPMA) have maintained an ongoing dialogue with major regulatory agencies. The ongoing dialogue is specifically referring to adaptive designs. For example, in preparation for the release of the EMA draft guideline, an expert group of pharmaceutical statisticians with an interest in adaptive designs and who were willing to share experiences met in September 2005 to identify potential opportunities for adaptive designs in late-phase clinical drug

development. The meeting was sponsored by PSI (Statisticians in the Pharmaceutical Industry), a professional association of statisticians interested in the application of statistics in the pharmaceutical industry. The article by Phillips and Keene [128] outlines the issues raised, resulting discussions and consensus views reached.

Similarly, in 2004, PhRMA had constituted a working group on adaptive clinical trial designs with the main objective to foster and facilitate wider consideration and understanding of adaptive designs and their potential for enhancing clinical development through fact-based evaluation of the benefits and challenges associated with these designs. Dragalin [129] provided an overview on terminology and classification in this area. In their 'Executive Summary' [3], the working group addressed the logistic, operational, procedural, and statistical challenges associated with adaptive designs. Three particular areas of adaptive design advantage have been emphasized: sample size re-estimation, dose finding, and seamless Phase II/III trials designs. A more extensive series of articles on these topics have been later published in the group's 'White Papers' [130]. A general definition of adaptive design is proposed as a multi-stage study design that uses accumulating data to decide on how to modify aspects of the study without undermining the validity and integrity of the trial. To maintain study validity means providing correct statistical inference, such as adjusted p-values, adjusted estimates, and adjusted confidence intervals. Trial validity is about credibility, interpretability, and persuasiveness of the study results. To maintain the study integrity means preplanning, as much as possible, based on intended adaptations, minimizing operational bias, maintaining the blind of interim analysis results, and assuring consistency between different stages of the study as a prerequisite for combining information from different trial stages. In short, the validity is about statistical bias, whereas integrity is about operational bias.

The ultimate objective of these industry-sponsored activities is to promote the acceptance of adaptive designs that are scientifically sound and defensible by all involved in the development and assessment of pharmaceutical products. This is particularly important considering the trend toward global trials to support simultaneous regulatory filings in all major geographic regions. Chuang-Stein *et al.* [131] summarized several occasions when members of these industry associations engaged with regulators from EMA, FDA, Health Canada, and the Pharmaceuticals and Medical Devices Agency in Japan (PMDA). These events drew support and expertise from many individuals involved in these discussions and reflect our common belief in the role adaptive designs could play in the pursuit of safe and efficacious medicines for human beings.

5. Applications

Adaptivity is a fundamentally important concept, which can be applied to all stages of drug discovery and development. To reap the most beneficial from adaptivity, an experimenter needs to carefully consider the reason for adaptation and possible adaptive choices. All adaptations have consequences, and it is important to articulate them when planning the experiment. Some adaptations may have broad applicability, others may only be occasionally useful, and some should probably be avoided altogether. Researchers should understand the distinctions among the various proposed adaptations and routinely consider the approaches that have shown promise in product research and development. In 2005, a literature search was performed to summarize the studies that were designed with the use of the combination testing principle or the conditional type I error rate approach [132]. The authors found about 60 papers and described basic element of the trial designs, which in general have been poorly described in methodological detail. Bretz *et al.* [133] moved through the drug discovery and development process and identified possible opportunities for adaptivity. Schmidli *et al.* [134] provided several examples of adaptive trial designs considered as feasible for confirmatory studies. In this section, we discuss three applications in more detail. The description is partially simplified but contains all essential features. We start with presenting a study with a sample size reassessment at an interim stage. In Section 5.2, we present a case study with an adaptive treatment selection in a two-stage design. Finally, in Section 5.3, we discuss adaptive endpoint selection.

5.1. Sample size reassessment

The following example illustrates the implementation of an adaptive group sequential design with sample size re-estimation in the Phase III clinical trial MUSEC (MUltiple Sclerosis and Extract of Cannabis, Trial Registration Number NCT00552604; for details see [135]), that investigated a standardized oral cannabis extract (CE) for the symptomatic relief of muscle stiffness and pain in adult patients with stable multiple

sclerosis and ongoing troublesome muscle stiffness. The primary outcome measure was an 11-point category rating scale measuring patient reported change in muscle stiffness from baseline to 12 weeks of treatment.

The pre-planned sample size calculations were based on the observed proportion of subjects with relief from muscle stiffness (0–3 categories on the category rating scale) in the CE and placebo arms in a previously conducted study on cannabinoids in multiple sclerosis: 0.42 and 0.27, respectively. A Fisher exact test for comparing such two proportions with 5% significance level and power 80% requires 170 evaluable subjects per arm. Adjusting for a dropout rate of 15%, the pre-planned total sample size was 400 subjects.

An unblinded interim analysis was planned after the first 200 subjects had completed the 12-week treatment. An early stopping for superiority using the O'Brien and Fleming boundary was considered as well as an unblinded sample size re-estimation procedure based on conditional power considerations for the second stage. The adjustment for these adaptations was implemented using the inverse normal p-value combination method with equal weights. At the time of the interim analysis, 101 subjects randomized to CE arm and 97 subjects to placebo had finished their 12-week treatment. The numbers of subjects with a relief from muscle stiffness in the CE and placebo arms were 27 and 12, respectively. The first stage one-sided p-value was 0.0055. Early rejection was almost reached considering the first stage adjusted significance level of the O'Brien-Fleming design being 0.0026. At the time of the interim analysis, 250 subjects had already been randomized, and the conditional power calculations (using the pre-planned or the observed effect as true effect) for a reduced total of 300 subjects still achieved values above 90%. Therefore, the Independent Data Monitoring Committee (IDMC) made the recommendation to reduce the patient number from 400 to 300. Note that by sticking to the original plan, the analogous conditional power calculations revealed values above 97%; hence, irrelevant effect differences in a large second stage would have caused a rejection at the final statistical analysis.

The study continued enrolling new subjects, and the final analysis was conducted when 143 subjects in the CE and 134 in the placebo arm completed their treatment. This was slightly below the planned target number. Overall, the rate of relief from muscle stiffness after 12 weeks was almost twice as high with CE than with placebo, 0.294 versus 0.157, the second stage rates were 0.357 versus 0.243. This yielded an inverse normal test statistic of 2.573 exceeding the critical boundary 1.977 of the final analysis. Hence, the difference was statistically significant.

An issue arises from decreasing rather than increasing the sample size. From the guidelines and current practice, a sample size increase seems to be appropriate, and a decrease is usually discouraged. However, if early rejection was considered a valid option in the first place, adding additional data in a second stage should allow both options – an increase but also a decrease of the planned total sample size. But we understand that statisticians must be rather brave to reduce the sample size because in case that a rejection cannot be reached in the final analysis, she or he could be blamed for having reduced the sample size.

The study shows that a decrease in sample size might be a reasonable option. The decrease was additionally justified by the fact that safety was not of a major concern so that there was no demand for a larger safety sample. From a company's perspective, the smaller necessary patient cohort seems to be attractive mainly because of the reduction in costs and time. Note also that this might be regarded as an alternative to adaptively adding interim analyses or preplanning group sequential trials with more stages from the beginning. In these cases, a similar reduction in patients might have been achieved, however, with the cost of doing more interim stages. We also note that the conditional error approach allows one to add an additional interim look, and so it would have been possible to stick to the original sample size 400 but to add an additional look at 300 patients. However, this was not foreseen in the study protocol and regulators required to prespecify the types of adaptation to be performed.

5.2. Treatment selection

The following example was the first major clinical trial using adaptive design methodology for many-to-one comparisons with dose-selection at an adaptive interim analysis. Zeymer *et al.* [136] conducted an international, prospective, randomized, double-blind, placebo-controlled Phase II trial in patients undergoing thrombolytic therapy or primary angioplasty for acute ST-elevation myocardial infarction applying a two-stage adaptive design. Fisher's combination function was pre-fixed to combine the adjusted p-values for all intersection hypotheses of a first stage trend test and individual dose control comparisons at the second stage ($\alpha = 0.025$, $\alpha_0 = 0.5$, $c = 0.0038$, and $\alpha_1 = 0.0101$). A binding futility boundary of $\alpha_0 = 0.5$ was proposed in the study protocol, meaning the trial would have to be stopped if the

dose response trend showed in the wrong direction. The primary efficacy end point was the immediate response variable infarct size measured by the cumulative release of α-HDBH within 72 h after administration of the drug (area under the curve, α-HDBH AUC). The objective of the adaptive interim analysis of the first stage was threefold: 'to obtain some initial evidence of the primary efficacy endpoint, to select a subset of doses to be carried forward into stage 2, and to determine the number of patient to be recruited for stage 2' [136].

The trial started with comparing four dose levels (50, 100, 150, and 200 mg eniporide) with placebo in 430 patients. In the first stage, the means $\pm$ SD (45.3 ± 31.8, 40.2 ± 22.5, 33.9 ± 20.5, 43.9 ± 27.0) were observed for the treatments groups (50, 100, 150, and 200 mg eniporide). In the placebo group, the first stage mean was 44.2 ± 26.0 resulting in p-value of $p = 0.12$ for the pre-specified one-sided linear trend test for the primary endpoint. As the largest effect (i.e., decrease of α-HDBH) was observed for the 150-mg dose, the interim look led to dropping the highest and the lowest dose group based on efficacy and safety arguments. Instead of using a trend test, a hierarchical test (starting with the highest selected dose versus placebo) was selected as multiplicity adjustment for the second stage. For the selected dose groups, a sample size recalculation yielded a sample size of 316 patients per group for the second stage. The authors used the conditional error of the trend test as local significance level for the sample size calculation of the second stage test. With 316 patients per group in the second stage conditionally on the already observed data, at the final analysis a power of 90% was targeted assuming a true difference between two groups of a quarter of the standard deviation and a local significance level of 0.032 ($= c/p$). In the second stage, 100 and 150 mg eniporide were compared with placebo in 959 patients. The positive findings of the first stage were not confirmed by the second stage data with observed means of (43.0 ± 26.1, 41.5 ± 25.9) for the two selected dose levels and 41.2 ± 28.5 for placebo. Combining first and second stage p-values by Fisher's combination function did not yield any statistically significant results for the global null hypothesis of interest. Therefore, no multiple testing of individual null hypotheses was performed. Although the planned multiple testing procedure described in the paper aims at keeping things simple (not addressing all aspects of the closed testing procedure), it can be seen as a first cautious attempt of a later called adaptive seamless design trying to combine the objectives of Phase II (Proof of Concept) and Phase III (testing of individual doses) into one trial. To summarize, the trial did not succeed in showing that the drug was superior to placebo at any of the investigated dose levels.

Other examples of adaptive treatment selection designs have been implemented since then [137]. INHANCE [138] was a multinational, multicenter, double blind, double dummy, two-stage adaptive, parallel group study design with blinded formoterol and open label tiotropium as active controls in patients with chronic obstructive pulmonary disease (COPD). This trial was planned as one of two pivotal trials to support registration and label claims of indacaterol as novel therapy for the treatment of COPD. The aim of the trial was to provide pivotal confirmation of efficacy, safety, and tolerability of the selected doses of indacaterol, where the dose selection is performed at a pre-specified interim analysis. In this case study, a two-stage Phase III adaptive design was an appropriate option, because 'dose' was the only major remaining question, and a large body of evidence was available at the end of Phase II. Overall, this design led to a reduction of approximately 15% in terms of development program length, number of patients, and costs as compared with a more traditional design of two sequential trials. INHANCE was included as a pivotal study in submissions to regulatory agencies globally, and indacaterol is now approved in all major markets globally for once-daily maintenance bronchodilator treatment of airflow obstruction in adult patients with COPD. The results of the interim analysis of INHANCE have been published in full [139], as have those of the final analysis [138].

This INHANCE study is particularly instructive to illustrate how operational biases can be minimized by using an IDMC to perform mid-trial adaptations based on predefined decision rules and review of unblinded results from an interim analysis. The IDMC was an autonomous group of recognized experts in the respiratory and statistical field. It was appointed by the study sponsor but functioned independently of all other persons involved with the study. The responsibilities of the IDMC in the INHANCE study included the following:

(1) revision and approval of the IDMC charter, which set out responsibilities, functions, rules of conduct, and the basis for evaluating the interim analysis results;
(2) review of the efficacy, safety, and tolerability results at the interim analysis; and
(3) selection of two indacaterol doses to be evaluated in the second stage of the trial, based on predefined criteria comparing indacaterol with placebo and the active controls.

An independent statistician (external to the sponsor) was appointed to produce the interim analysis report and interact with the IDMC. The independent statistician had access to the clinical data and the treatment codes (labeled from 'A' to 'G') but not the actual treatment descriptions. In addition, a communication plan between the IDMC and the sponsor was included in the IDMC charter. According to this plan, if there were no complexities in the data and the IDMC did not see any reason to deviate from the predefined dose selection guidelines, the IDMC chair was to inform the sponsor of the recommended doses only. That is, no quantitative results were released to minimize the possibility for adaptations to convey information to observers about the interim results and mid-trial changes [140]. In the INHANCE study, the IDMC did not have any safety concerns and selected two doses to continue into the second stage of the study, based on the guidelines specified in the IDMC charter and without any discussion with the sponsor. Therefore, no one from the sponsor was aware of the interim results before final database lock. More details on the methodology and the logistics implemented in the INHANCE study can be found in [141, 142].

Hemangeol was developed as the treatment for proliferating infantile hemangioma requiring systemic therapy. It was developed for the use in pediatric population following the guidelines of health regulatory agencies. The Phase III development of Hemangeol was based on a two-stage adaptive confirmatory trial with regimen selection at the end of the first stage, in order to identify the appropriate dose and duration for further study in the second stage. Early stopping for futility and sample size re-estimation were also considered at the interim analysis. The aim of this trial was to demonstrate the superiority of the selected dose(s) over placebo and to establish its safety profile. Marketing authorization of Hemangeol was granted by both FDA and EMA in 2014. This case study highlights operational challenges of adaptive designs as by the time of the pre-planned interim analysis recruitment targets, and thus the necessary sample size had been reached. It was decided before unblinding, however, to perform the interim analysis as planned so that recruitment could continue without any pauses, should sample size reassessment be necessary. Leaute-Labrze *et al.* [143] published full details about the trial design and analysis, together with all related documents, including the clinical study protocol. Heritier *et al.* [144, 145] summarized their practical experiences when conducting this two-stage adaptive confirmatory trial, focusing on statistical aspects, logistical challenges and regulatory interactions.

Secretory diarrhea in HIV positive patients remains a serious unmet clinical need, even and especially in the age of highly active anti-retroviral therapy. In late 2012, crofelemer was approved by the FDA as a first-in-class anti-diarrheal agent indicated for the symptomatic relief of non-infectious diarrhea in adult HIV patients on anti-retroviral therapy. The safety and efficacy of crofelemer were established through ADVENT, a two-stage adaptive clinical trial with dose selection at the end of the first stage [146].

5.3. Endpoint selection

A placebo-controlled two-arm multicenter trial was performed to test three co-primary endpoints – two superiority and one non-inferiority hypotheses have been involved. It was investigated whether clonidine as a co-medication with fentanyl and midazolam is superior to fentanyl and midazolam alone in ventilated newborns, and infants up to 2 years of age, as measured by the endpoints: total consumption of fentanyl (superiority hypothesis 1) and midazolam (superiority hypothesis 2). Additionally, non-inferiority to placebo with respect to the need for the rescue thiopentone use has to be shown (hypothesis 3). It was a study of the PAED-Net, which is a corporation of pediatric modules in six German university locations in a small and vulnerable population. For details of the study, conduct see [147] (Trial Registration Number ISRCTN77772144).

Regulatory authorities usually expect statistical significance in all three co-primary endpoints simultaneously, which would mean that no further multiple testing correction is needed. But in the setting of pediatric populations the investigators were convinced that under double blind conditions it would be worth to achieve significance in at least one of the two superiority hypotheses. The noninferiority margin for thiopentone use was set to 20%. The corresponding global null hypothesis was tested using the OLS test, and a closed testing was planned in order to show significant differences in specific endpoints [148–150]. The study was planned in three stages with critical values according to O'Brien and Fleming adjusted significance levels 0.0003, 0.0071, and 0.0225, respectively. The results of the three stages were combined using the inverse normal method together with the closed testing principle as described in Section 3.4. Under reasonable assumption of the effect sizes and taking into account the correlation of the endpoints, a sample size of 210 patients was estimated to achieve 80% overall power at one-sided significance level $\alpha = 0.025$. Overall power was defined for detecting at least one significant difference. Considering the power to reject all hypotheses required a lot more patients and was considered inappropriate.

Three types of adaptations were pre-specified in the study protocol:

(1) a sample size reassessment based on the results observed at interim stages,
(2) the possibility to redefine the test statistic if it turned out that the OLS test statistic was clearly inferior to a better overall test,
(3) dropping a superiority endpoint if the effect size in this endpoint was too low.

The last option seemed to be useful because it was not clear at the beginning if both fentanyl and midazolam consumption could be reduced with clonidine.

The first interim analysis yielded very promising results: The overall p-value for the OLS test was 0.0009, thus already near showing significance. The midazolam result, however, was already weak ($p = 0.0472$) as compared with the others ($p = 0.0051$ for fentanyl and $p = 0.0012$ for thiopentone). This trend was dramatically confirmed at the second interim analysis: A negative effect in midazolam was observed for the second stage data, yielding a p-value of 0.678. Nevertheless, the OLS test for the global multivariate hypothesis yielded a p-value of 0.0017. This was statistically significant, although a study continuation was recommended because superiority with regard to fentanyl was not very clear anymore. Particularly, it was not significant within the closed test procedure, only noninferiority with regard to thiopentone was significant within the closed test procedure. The study recommendation was also to drop midazolam consumption as a primary endpoint from the further analysis because it might jeopardize an overall positive result of the study.

Although a positive result was likely for the reduced clinical question, at the final analysis, even fentanyl could not be shown to be significant. The main study results were published in [147].

The example shows the importance of keeping adaptive interim decisions secretly. If the doctors had been aware of the fact that midazolam was dropped from further confirmatory analyses, this might have clearly influenced treatment and medication of the patients. In order to exclude this possibility, only the Independent Statistical Center and the IDMC who gave the recommendation were aware of the study results. The head of the study was informed that there was a recommendation, the decision on it was left to one representative of the sponsor (Boehringer Ingelheim, Germany) that needed to be involved. This again illustrates that trial logistics is an important issue in adaptive designs. It needs to be extensively discussed how operational biases can occur and how they can be avoided.

Although the study did not show the desired effect, especially the second interim analysis illustrates the potential advantage of an adaptive way of analyzing data. There were no strict stopping criteria, and the continuation of the trial produced a disappointing result but obviously reflects reality. The study was planned in 2002 and finalized in 2008. Publication of the non-convincing study results was a problematic issue. Another issue with publishing complicated adaptive designs in medical journals is to have space to communicate the statistical methodology [132]. Several details of the statistical study design were not provided in [147]. We nevertheless think that this study serves as an interesting example for an early attempt for an adaptation, which goes beyond sample size reassessment and treatment arm selection in a vulnerable, small population [151].

6. Software

Nowadays, the availability of software is a necessary condition for the applicability and acceptance of a statistical methodology. Many of the procedures proposed for adaptive designs additionally require high-computational effort such that software should be able to perform time consuming computations in a relatively short time. Up to now, the reviews of software packages on clinical trials with interim analysis concentrated on packages specifically designed for group sequential methods [152–154], the reason simply being that software for adaptive designs was not available at that time. One review of software for adaptive designs [155] appeared recently.

There is wide field of available commercial software for group sequential designs (e.g., ADDPLAN, East, SEQTEST SEQDESIGN from SAS, nQuery's nTerim, PASS from NCSS) and R-packages [156–160], which can be used to determine critical boundaries for confirmatory adaptive designs that are based on the inverse normal combination test. Christopher Jennison provides Fortran programs for group sequential designs freely available on his homepage: www.bath.ac.uk\ ~mascj. Further, Fortran programs for the computation of the use function approach are available from the University of Wisconsin School of Medicine and Public Health site www.medsch.wisc.edu\Software\landemets\: Programs for Computing Group Sequential Boundaries Using the Lan-DeMets Method, Version 2.1.

Specifically for confirmatory adaptive designs, there is still only a limited number of available software, both commercially and non-commercially. Since the very beginning of adaptive designs, the software ADDPLAN was created for adaptive confirmatory designs (www.addplan.com). It is commercially available since 2002 as a stand alone tool for designing, simulating and performing analysis with an emphasis on the adaptive confirmatory technique. The MC module provides additional multiple comparison features for more than two treatment arms in simulation and analysis, and the PE module additional features for patient enrichment designs in simulation and analysis. There is also the new DF module with capabilities for adaptive dose-finding designs.

East from Cytel (www.cytel.com) is a comprehensive tool for design, simulation and analysis of trials with interim analyses. Since 2006, adaptive extensions are provided with the East ADAPT and the East SURVADAPT module. Recently, the modules East MULTIARM and East ENDPOINT provide extensions to multi-arm designs and designs with multiple endpoints. An extension to dose finding trials comes with East ESCALATE and Cytel's COMPASS.

We checked CRAN (Comprehensive R Archive Network) cran.rstudio.com on Jan 20, 2015, for R-packages available for confirmatory adaptive designs. We list the packages with a short description. It is emphasized that this is a dynamic development, and we expect many more packages in the near future.

- adaptTest: Adaptive two-stage tests [161]. The functions defined in this program serve for implementing adaptive two-stage adaptive tests that are based on the combination testing principle.
- AGSDest: Estimation in adaptive group sequential trials [162]. This module enables the calculation of confidence intervals in adaptive group sequential trials.
- asd: Simulations for adaptive seamless designs [163]. This package runs simulations for adaptive seamless designs with and without early outcomes for treatment selection and population enrichment type designs.
- gMCP [164]: This program provides functions and a GUI for adaptive [91] and non-adaptive [86] graph-based multiple comparison procedures.
- interAdapt [165]. This is an interactive tool for designing and evaluating certain types of adaptive enrichment designs.

There is also an R-package called RCTDesign: Methods and Software for Clinical Trials, which is freely available at www.rctdesign.org. This package builds on the formerly available S-Plus module S+SeqTrial and has already an add-on for adaptive methods. Furthermore, the book [166] contains some R-programs for adaptive designs. It also includes programs for performing sample size reassessment procedures and some basic adaptive randomization designs. The book also comes with SAS macros, the most of them performing simulations for the adaptive designs described in the book. Further, SAS macros based on SAS IML have been developed for many-to-one comparisons with adaptive treatment selection [60].

To summarize, some software is free and hence attractive for statistical research. This is particularly true for the increasing number of available R-packages. Simulation-based evaluation of operating characteristics of adaptive designs is becoming increasingly important, some of the available adaptive R-packages already include such functionality. Within commercially available packages, only ADDPLAN, East ADAPT and East SURVADAPT address the specific requirements for confirmatory adaptive designs.

7. Summary and discussion

Adaptive confirmatory designs have shaken the classical design paradigm that the details of the design and statistical analysis all have to be all laid down in advance. There are two ways on how to achieve the flexibility of mid-trial design modifications, which in principle need not to pre-specified. In the first approach, separate test statistics (such as p -values or z-values) are calculated from the disjoint samples at the different stages and combined into an overall test statistics in a fixed pre-defined way. This approach relies on invariance properties of the stagewise test statistics, for example, that the p -values under the null hypothesis are uniformly distributed. In the second approach, the probability of an erroneous rejection of the null hypothesis by the pre-planned design given the data at interim (conditional error rate) is calculated at the interim analysis. The remainder of the pre-planned design then can be replaced by any other design, which never would raise a larger conditional error rate. Both methods allow a frightening multitude of flexibility, some of it having been sketched in this review. However, in case adaptations are made by leaving the setting of group sequential designs, both approaches use test statistics that are different from the minimal sufficient test statistics of conventional tests. Curious examples can be constructed that

extremely different weighting of sample units from the different stages can produce absurd results. One way to avoid such scenarios is to plan for some reasonable constraints on flexibility or to allow a rejection of the null hypothesis only if also the conventional test statistics would reject, for example, at full level α. Clearly, this latter 'dual test' comes at the cost of a slight decrease in power but avoids borrowing strength from a curious weighting of the stagewise test statistics [55, 56, 167, 168]. Whenever the conditional error of the adapted design (using the conventional analysis) is lower than the one of the originally planned design, the dual test is implicitly applied. This property is used, for example, by Mehta *et al.* [169] when increasing the sample size for 'promising' interim effects [54] still using a conventional analysis; see also a comment by Glimm [170].

Note that the well-founded sufficiency principle might obviously be considered as inappropriate or at least too strongly formulated in the context of providing a statistical methodology for clinical research. The choice of a methodology is not only guided by mathematical optimality but also by practical demands, and the price for adaptivity is the concession to use a suboptimal test statistic in case an adaptation was performed. Interestingly, the Bayesian principle of combining prior beliefs with new evidence from data and thereby allowing adaptivity *by definition* might be better suited for the practical demands but lacks type I error rate control, at least in general. Nevertheless, Bayesian methods [73] can be used for assessing interim results (e.g., predictive power), and thus the methods can be used within the available methodology for adaptive designs for confirmatory trials with frequentist analysis.

If hypotheses are adaptively changed during an ongoing trial, multiple testing adjustments will be needed in order to get valid inference on the hypotheses under investigation, allowing conclusive interpretation of the study results. This is a basic requirement of scientific rigor not to bypass statistical principles by abusing flexibility in a wrong direction. It is not surprising that multiple testing in connection with adaptive choice of hypotheses introduce a further level of methodological complexity. Also, estimation is not a simple problem in adaptive design. A lot of research over the last years have tackled these issues and has given us a better understanding of inference in flexible designs. But obviously, there is another challenge that arises from how to apply adaptive designs in the environment of clinical trials. It took quite a time till (group) sequential designs have found their way into clinical trial routine. Nowadays, interim analyses, at least to decide whether to stop for futility, have become an important feature of (large) clinical trials to account for ethics and costs. Clearly, the logistics and workload for unblinded adaptive interim analyses (who is doing it, what information is passed on, who is getting informed and who is deciding on adaptations?) are more demanding than for interim analyses in conventional group sequential designs. Hence, the required input may be prohibitive to consider adaptive designs as an option in several standard clinical trial scenarios.

The adaptive approach was essentially laid down by the seminal papers in the 90's of the last century. It was accomplished by some kind of natural skepticism from the methodological and the regulatory perspective. In the meantime, there is some clarification of how and when to use adaptive designs. Some concern is still being caused by the unblinding of study results at interim stages. This is an essential feature of classical group sequential designs, and hence part of criticism on adaptive designs is inherited from the latter approach. However, adaptation must not necessarily be based on breaking the blind. One of the simplest adaptive designs is the blinded SSR design that usually consists of two stages in which the sample size for the second stage is determined based on the first-stage data. This was introduced as the 'internal pilot design' at more or less the same time [8, 9]. The blinded SSR design determines the sample size of the second stage using only the estimate of nuisance parameters such as variance or overall standard deviation, overall response rate, or overall survival pattern (see, e.g., [171–174]). This design is easy to implement and generally does not require adjustments. It is quite efficient if the true treatment effect is close to the pre-planned target that is fixed in such designs. However, the pre-planned target treatment effect is often based on an optimistic guess of the 'clinically relevant' treatment difference at which the power is specified. We think that even for the unblinded cases, a careful consideration of both the effect size and the variability should serve as a guideline for making interim decisions including SSR.

There is no such thing as a free lunch. Therefore, in planning a study, it should be carefully checked in advance, whether in the specific situation the existing caveats are dominated by the potential benefit of flexibility. It has to be kept in mind, however, that in a specific scenario, a method allowing flexibility of going beyond the specific scenario will not be able to beat a method that is optimally tailored to just that one scenario. The advantage of the group sequential version of adaptive designs is that if no adaptations are made at interim, the 'optimally' planned design will be performed without paying any price for the option of flexibility. The price would be paid if the adaptation detracts the experimenter from the truly optimal design. Note, however, that all these different situations are rather theoretical and can hardly be

verified in real-life trials. The fundamental pros and cons have been exchanged. Some early enthusiasm from clinical trialists who hoped to get new useful tools may also have been caused by not fully understanding the inferential complications created by flexibility. This may also have been a motivation of regulators in their guidance documents to set rather narrow limits for applying unblinded adaptive designs in drug development (e.g., with regard to the involvement of the sponsor in the interim analysis). Hence, as for many other research dealing with innovative treatments of patients naïve promises arising at the beginning here and there did not become real in the present clinical trial community. There are successful examples of carefully planning and running adaptive clinical trials, which the proponents considered to have been helpful in drug development. So the proof of principle on how to use the methodology for different types of adaptations has been given. By the way, there are also other areas like genomics with huge number of hypotheses where some basic ideas of the adaptive testing principle have been taken over.

It should also be kept in mind that adaptive designs are a very potent tool to deal with the unexpected in clinical trials. Not all the details necessary for optimally designing a clinical trial are available in the planning phase. If they were available rather, no trial would be planned at all. Whenever serious deviations from the planning assumptions become obvious at interim, it may be advantageous as an 'ultima ratio' to use the adaptive design methodology for overcoming deficiencies and 'saving' the running trial, for example, by applying the conditional error rate approach. This would be a reasonable way to deal with protocol amendments containing substantial design modifications in a scientifically honest way. Protocol amendments in practice are an extensively used tool often to open a short cut for introducing flexibility in ongoing clinical trials, which is not accounted for in the statistical analysis.

Beyond all skepticism, there are opportunities where adaptive design serve as a valuable tool and generalization of commonly used methods. In many of the cases where (unblinded) interim analyses are performed, there is the potential to extend these to include data driven changes of the design. Clearly they cannot be applied and even do not help for all cases. For example, maintaining type I error rate control using all interim information might be critical. In cases where patients for further stages were already recruited and data which is related to the endpoint is already available at an earlier stage, type I error rate control is violated. This is specifically the case in survival designs. As briefly reviewed here, providing solutions is the topic of current research. Although there are methodological caveats, the incorporation of adaptive elements in study designs seems to work in general. However, sometimes, it is even not preferable to incorporate interim results, especially from small stages, for example, because information from interim results might be insecure, and it might be better to stick to the original assumptions rather than trusting the interim results [57, 175], or a treatment arm or population is too often wrongly selected and therefore yields power smaller than a conventional design. This is not surprising but often overseen when people want to look at the data as soon as possible.

Designing an adaptive design is a complex task and more statistical planning is needed to account for the flexibility introduced in the study design. Specifically, statistical software is increasingly needed to evaluate the adaptations and to find reasonable strategies. The question might arise if potential decisions made at interim stages might not be better placed to the upfront planning stage.

Adaptive designs also come along with more operational and organizational problems. If the sponsor is not able or not willing to recruit an additional number of patients, it makes no sense to ask for sample size increase. In this case, stopping for futility might be the only option that is worthwhile to consider. If the sponsor agrees to a potential sample size increase, the randomization process including drug supply needs to be managed. Is it possible to correctly administer the drug according to a selected dose regimen? How does the concrete adaptation work? If an IDMC is doing the interim analysis, the IDMC can only recommend a design change, its decision is still endorsed by a sponsor's representative. This makes the dissemination of study results an issue and the careful formulation of the information flow, for example, in an IDMC charter, becomes an essential part of designing the trial. Some feasible models for implementation of adaptive designs have been proposed by the industry [176–178].

In summary, adaptive designs have been carefully developed in the past 25 years and – at least from a theoretical perspective – their properties, advantages and disadvantages, are well understood. To achieve full acceptance in the statistical community and by regulators, there is still the need for both more methodological expertise [179] and practical experience. The more this exciting methodology will be used the more it will be understood when it is helpful and when it is not.

Acknowledgements

We thank the editor Ralph D'Agostino for the invitation to write this review. We are grateful to Stephane Heritier and five anonymous referees for their helpful comments, which improved the presentation of the article. This project has received funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration under grant agreement no 602552 (IDEAL - Integrated Design and Analysis of small population group trials).

References

1. Bauer P. Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie* 1989; **20**:130–148.
2. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041, correction in *Biometrics* 1996, **52**:380.
3. Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug development – an executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics* 2006; **16**:275–283.
4. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
5. Elsäßer A, Regnstrom J, Vetter T, Koenig F, Hemmings RJ, Greco M, Papaluca-Amati M, Posch M. Adaptive clinical trial designs for European marketing authorization: a survey of scientific advice letters from the European Medicines Agency. *Trials* 2014; **15**(1):383.
6. EMA. *Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design*. European Medicines Agency: London, UK, 2007. http://www.ema.europa.eu.
7. FDA. *Draft Guidance for Industry Adaptive Design Clinical Trials for Drugs and Biologics*. Food and Drug Administration. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER): Rockville MD, USA, 2010. url=http://http://www.fda.gov/.
8. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* 1990; **9**:65–72.
9. Birkett MA, Day SJ. Internal pilot studies for estimating sample size. *Statistics in Medicine* 1994; **13**:2455–2463.
10. Friede T, Kieser M. Sample size recalculation in internal pilot study designs: a review. *Biometrical Journal* 2006; **48**(4):537–555.
11. Posch M, Proschan MA. Unplanned adaptations before breaking the blind. *Statistics in Medicine* 2012; **31**(30): 4146–4153.
12. Schneider S, Schmidli H, Friede T. Blinded sample size re-estimation for recurrent event data with time trends. *Statistics in Medicine* 2013; **32**(30):5448–5457.
13. Proschan M, Glimm E, Posch M. Connections between permutation and t-tests: relevance to adaptive methods. *Statistics in Medicine* 2014; **33**(27):4734–4742.
14. Tsiatis AA, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**: 367–378.
15. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trial. *Statistics in Medicine* 2003; **22**:971–993.
16. Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine* 2006; **25**(6):917–932.
17. Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. *Biometrika* 2006; **93**:1–21.
18. 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**(6):1956–1961.
19. Burman C-F, Sonesson C. Are flexible designs sound? *Biometrics* 2006; **62**(3):664–669.
20. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**(4): 1286–1290.
21. Cui L, Hung HMJ, Wang S-J. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**: 853–857.
22. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**: 1315–1324.
23. 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.
24. 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.
25. Berry SM, Carlin BP, Lee JJ, Muller P. *Bayesian Adaptive Methods for Clinical Trials*. CRC Press: Boca Raton, 2010.
26. Hu F, Rosenberger WF. *The Theory of Response-Adaptive Randomization in Clinical trials*. John Wiley & Sons: Hoboken, 2006.
27. Case LD, Morgan TM, Davis CE. Optimal restricted two-stage designs. *Controlled Clinical Trials* 1987; **8**:146–156.
28. Gugerli US, Maurer W, Mellein B. Internally adaptive designs for parallel group trials. *Drug Information Journal* 1993; **27**:721–732.
29. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 2003; **22**(5):689–703.
30. Graf AC, Bauer P, Glimm E, Koenig F. Maximum type 1 error rate inflation in multiarmed clinical trials with adaptive interim sample size modifications. *Biometrical Journal* 2014; **56**(4):614–630.
31. Mehta CR, Tsiatis AA. Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal* 2001; **35**:1095–1112.

32. Zhou J, Adewale A, Shentu Y, Liu J, Anderson K. Information-based sample size re-estimation in group sequential design for longitudinal trials. *Statistics in Medicine* 2014; **33**(22):3801–3814.
33. Posch M, Maurer W, Bretz F. Type I error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim. *Pharmaceutical Statistics* 2011; **10**(2):96–104.
34. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Academic Press: New York, 1985.
35. Bauer P. Sequential tests of hypotheses in consecutive trials. *Biometrical Journal* 1989; **31**:663–676.
36. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statistical Association* 2002; **97**:236–244.
37. Fisher LD. Self-designing clinical trials. *Statistics in Medicine* 1998; **17**:1551–1562.
38. Hartung J. A self-designing rule for clinical trials with arbitrary variables. *Controlled Clinical Trials* 2001; **22**: 111–116.
39. Wassmer G. Multistage adaptive test procedures based on Fisher's product criterion. *Biometrical Journal* 1999; **41**: 279–293.
40. Posch M, Bauer P. Adaptive two stage designs and the conditional error function. *Biometrical Journal* 1999; **41**:689–696.
41. Wassmer G. *Statistische Testverfahren für gruppensequentielle und adaptive Pläne in klinischen Studien. Theoretische Konzepte und deren praktische Umsetzung mit SAS*. Verlag Alexander Mönch: Köln, 1999.
42. Wassmer G. A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics* 1998; **54**:696–705.
43. Brannath W, Bauer P. Optimal conditional error functions for the control of conditional power. *Biometrics* 2004; **60**: 715–723.
44. Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal* 1999; **41**:261–277.
45. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* 2001; **43**:581–589.
46. Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association* 2002; **97**:1034–1041.
47. Liu Q, Pledger GW. On design and inference for two-stage adaptive clinical trials with dependent data. *Journal of Statistical Planning and Inference* 2006; **136**:1962–1984.
48. Brannath W, Gutjahr G, Bauer P. Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association* 2012; **107**:824–832.
49. Posch M, Timmesfeld N, König F, Müller H-H. Conditional rejection probabilities of Student's t-test and design adaptation. *Biometrical Journal* 2004; **46**:389–403.
50. Timmesfeld N, Schäfer H, Müller H-H. Increasing the sample size during clinical trials with t-distributed test statistics without inflating the type I error rate. *Statistics in Medicine* 2007; **26**(12):2449–2464.
51. Gutjahr G, Brannath W, Bauer P. An approach to the conditional error rate principle with nuisance parameters. *Biometrics* 2011; **67**(3):1039–1046.
52. Shun Z, Yuan W, Brady WE, Hsu H. Type I error in sample size re-estimations based on observed treatment difference. *Statistics in Medicine* 2001; **20**(4):497–513.
53. Friede T, Kieser M. A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine* 2001; **20**(24):3861–3873.
54. Chen YH, DeMets DL, Lan KKG. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine* 2004; **23**(7):1023–1038.
55. Denne JS. Sample size recalculation using conditional power. *Statistics in Medicine* 2001; **20**:2645–2660.
56. Posch M, Bauer P, Brannath W. Issues in designing flexible trials. *Statistics in Medicine* 2003; **22**(6):953–969.
57. Bauer P, König F. The reassessment of trial perspectives from interim data – a critical view. *Statistics In Medicine* 2006; **25**:23–36.
58. ICH. *Topic E 9 Statistical Principles for Clinical Trials*, 1998. http://www.ema.europa.eu.
59. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
60. Bretz F, König F, Brannath W, Glimm E, Posch M. Tutorial in biostatistics: Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* 2009; **28**:1181–1217.
61. Lehmacher W, Kieser M, Hothorn L. Sequential and multiple testing for dose-response analysis. *Drug Information Journal* 2000; **34**:591–597.
62. Posch M, König F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimating in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* 2005; **2443**:3697–3714.
63. Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal* 2006; **48**(4):623–634.
64. Jennison C, Turnbull BW. Adaptive seamless designs: selection and prospective testing of hypotheses. *Journal of Biopharmaceutical Statistics* 2007; **17**(6):1135–1161.
65. Maca J, Bhattacharya S, Dragalin V, Gallo P, Krams M. Adaptive seamless phase II/III designs – background, operational aspects, and examples. *Drug Information Journal* 2006; **40**:463–473.
66. Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal* 2006; **48**(4):635–643.
67. Wassmer G. On sample size determination in multi-armed confirmatory adaptive designs. *Journal of Biopharmaceutical Statistics* 2011; **21**:802–817.
68. Friede Tim, Stallard Nigel. A comparison of methods for adaptive treatment selection. *Biometrical Journal* 2008; **50**(5):767–781.
69. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.
70. Magirr D, Stallard N, Jaki T. Flexible sequential designs for multi-arm clinical trials. *Statistics in Medicine* 2014; **33**: 3269–3279.

71. Gao P, Liu L, Mehta C. Adaptive sequential testing for multiple comparisons. *Journal of Biopharmaceutical Statistics* 2014; **24**(5):1035–1058.
72. König F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. *Statistics in Medicine* 2008; **27**:1612–1625.
73. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, Racine-Poon A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy on oncology. *Statistics in Medicine* 2009; **28**:1445–1463.
74. Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* 2011; **10**:347–356.
75. Wang S-J, O'Neill RT, Hung HM. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics* 2007; **6**(3):227–244.
76. Wang S-J, Hung HMJ, O'Neill RT. Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal* 2009; **51**(2):358–374.
77. Graf A, Posch M, Koenig F. Adaptive designs for subpopulation analysis optimizing utility functions. *Biometrical Journal* 2015; **57**(1):76–89.
78. Wassmer G, Dragalin V. Designing issues in confirmatory adaptive population enrichment trials. *Journal of Biopharmaceutical Statistics* 2015; **25**:early view.
79. Mehta C, Gao P, Bhatt DL, Harrington RA, Skerjanec S, Ware JH. Optimizing trial design sequential, adaptive, and enrichment strategies. *Circulation* 2009; **119**(4):597–605.
80. Mehta C, Gao P. Population enrichment designs: case study of a large multinational trial. *Journal of Biopharmaceutical Statistics* 2011; **21**(4):831–845.
81. Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine* 2012; **31**(30):4309–4320.
82. Stallard N, Hamborg T, Parsons N, Friede T. Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of Biopharmaceutical Statistics* 2014; **24**(1):168–187.
83. O'Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials* 1997; **18**:550–556.
84. Hung HMJ, Wang S-J, O'Neill R. Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *Journal of Biopharmaceutical Statistics* 2007; **17**(6):1201–1210.
85. Tamhane AC, Wu Y, Mehta CR. Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (ii): sample size re-estimation. *Statistics in Medicine* 2012; **31**(19):2041–2054.
86. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; **28**(4):586–604.
87. Bretz F, Posch M, Glimm E, Klinglmüller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal* 2011; **53**(6):894–913.
88. Burman C-F, Sonesson C, Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine* 2009; **28**(5):739–761.
89. Sugitani T, Hamasaki T, Hamada C. Partition testing in confirmatory adaptive designs with structured objectives. *Biometrical Journal* 2013; **55**(3):341–359.
90. Sugitani T, Bretz F, Maurer W. A simple and flexible graphical approach for adaptive group-sequential clinical trials. *Journal of Biopharmaceutical Statistics* 2014; **Accepted**.
91. Klinglmueller F, Posch M, Koenig F. Adaptive graph-based multiple testing procedures. *Pharmaceutical Statistics* 2014; **13**(6):345–356.
92. Schäfer H, Müller H-H. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine* 2001; **20**:3741–3751.
93. Wassmer G. Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal* 2006; **48**(4):714–729.
94. Desseaux K, Porcher R. Flexible two-stage design with sample size reassessment for survival trials. *Statistics in Medicine* 2007; **26**(27):5002–5013.
95. Jahn-Eimermacher A, Ingel K. Adaptive trial design: A general methodology for censored time to event data. *Contemporary Clinical Trials* 2009; **30**(2):171–177.
96. Bauer P, Posch M. Letter to the Editor: Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine* 2004; **23**:1333–1335.
97. Mehta C, Schäfer H, Daniel H, Irle S. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statistics in Medicine* 2014; **33**(26):4515–4531.
98. Irle S, Schäfer H. Interim design modificcations in time-to-event studies. *Journal of the American Statistical Association* 2014; **107**:341–348.
99. Magirr D, Jaki T, Koenig F, Posch M. Adaptive survival trials. *arXiv preprint arXiv:1405.1569* 2014.
100. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine* 2010; **29**(9):959–971.
101. Friede T, Parsons N, Stallard N, Todd S, Valdes Marquez E, Chataway J, Nicholas R. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine* 2011; **30**(13):1528–1540.
102. Carreras M, Gutjahr G, Brannath W. Adaptive seamless designs with interim treatment selection: a case study in oncology. *Statistics in Medicine* 2015; **Early View**.
103. Faldum A, Hommel G. Strategies for including patients recruited during interim analysis of clinical trials. *Journal of Biopharmaceutical Statistics* 2007; **17**(6):1211–1225.
104. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC: Boca Raton, London, New York, Washington, D.C., 2000.
105. Proschan M, Lan G, Wittes J. *Statistical Monitoring of Clinical Trials. A Unified Approach*. Springer: Science & Business Media, 2006.

106. Lawrence J, Hung HM. Estimation and confidence intervals after adjusting the maximum information. *Biometrical Journal* 2003; **45**(2):143–152.
107. Cheng Y, Shen Y. Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics* 2004; **60**(4):910–918.
108. Brannath W, König F, Bauer P. Estimation in flexible two stage designs. *Statistics in Medicine* 2006; **25**(19):3366–3381.
109. Coburger S, Wassmer G. Sample size reassessment in adaptive clinical trials using a bias corrected estimate. *Biometrical Journal* 2003; **45**(7):812–825.
110. Coburger S, Wassmer G. Conditional point estimation in adaptive group sequential test designs. *Biometrical Journal* 2001; **43**(7):821–833.
111. Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* 2008; **50**(4):515–527.
112. Bowden J, Glimm E. Conditionally unbiased and near unbiased estimation of the selected treatment mean for multistage drop-the-losers trials. *Biometrical Journal* 2014; **56**(2):332–349.
113. Mehta CR, Bauer P, Posch M, Brannath W. Repeated confidence intervals for adaptive group sequential trials. *Statistics in Medicine* 2007; **26**(30):5422–5433.
114. Brannath W, Mehta CR, Posch M. Exact confidence bounds following adaptive group sequential tests. *Biometrics* 2009; **65**(2):539–546.
115. Gao P, Liu L, Mehta C. Exact inference for adaptive group sequential designs. *Statistics in Medicine* 2013; **32**(23): 3991–4005.
116. Magirr D, Jaki T, Posch M, Klinglmüller F. Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika* 2013; **100**(4):985–996.
117. Bauer P, Koenig F, Brannath W, Posch M. Selection and bias – two hostile brothers. *Statistics in Medicine* 2010; **29**(1): 1–13.
118. Goll A, Bauer P. Two-stage designs applying methods differing in costs. *Bioinformatics* 2007; **23**(12):1519–1526.
119. Scherag A, Hebebrand J, Schäfer H, Müller H-H. Flexible designs for genomewide association studies. *Biometrics* 2009; **65**(3):815–821.
120. Victor A, Hommel G. Combining adaptive designs with control of the false discovery rate–a generalized definition for a global p-value. *Biometrical Journal* 2007; **49**(1):94–106.
121. FDA. *Innovation or stagnation: Critical path opportunities list*, U.S. Food and Drug Administration, 2006. http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPath Initiative/CriticalPathOpportunitiesReports/ UCM077258.pdf.
122. Hung HMJ, Wang S-J, Yang P. Some challenges with statistical inference in adaptive designs. *Journal of Biopharmaceutical Statistics* 2014; **24**(5):1059–1072.
123. Chen JJ, Lu T-P, Chen D-T, Wang S-J. Biomarker adaptive designs in clinical trials. *Translational Cancer Research* 2014; **3**(3):279–292.
124. Wang S-J, Hung HM. Adaptive enrichment with subpopulation selection at interim: Methodologies, applications and design considerations. *Contemporary Clinical Trials* 2013; **36**(2):673–681.
125. Wang S-J, Brannath W, Brückner M, James Hung HM, Koch A. Unblinded adaptive statistical information design based on clinical endpoint or biomarker. *Statistics in Biopharmaceutical Research* 2013; **5**(4):293–310.
126. Wang S-J, Hung HM, O'Neil RT. Impacts on type I error rate with inappropriate use of learn and confirm in confirmatory adaptive design trials. *Biometrical Journal* 2010; **52**(6):798–810.
127. Eichler H-G, Bloechl-Daum B, Abadie E, Barnett D, König F, Pearson S. Relative efficacy of drugs: an emerging issue between regulatory agencies and third-party payers. *Nature Reviews Drug Discovery* 2010; **9**(4):277–291.
128. Phillips AJ, Keene ON. Adaptive designs for pivotal trials: discussion points from the PSI adaptive design expert group. *Pharmaceutical Statistics* 2006; **5**(1):61–66.
129. Dragalin V. Adaptive designs: Terminology and classification. *Drug Information Journal* 2006; **40**:425–435.
130. PhRMA Working Group on Adaptive Designs. White papers. *Drug Information Journal* 2006; **40**(22):421–484.
131. Chuang-Stein C, Bretz F, Komiyama O, Quinlan J. Interactions with regulatory agencies to enhance the understanding and acceptance of adaptive designs. *Regulatory Affairs Focus* 2009; **14**(4):36–42.
132. Bauer J, Einfalt P. Application of adaptive designs - a review. *Biometrical Journal* 2006; **8**:1–16.
133. Bretz F, Branson M, Burman CF, Chuang-Stein C, Coffey CS. Adaptivity in drug discovery and development. *Drug Development Research* 2009; **70**:169–190.
134. Schmidli H, Bretz F, Racine-Poon A. Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. *Statistics in Medicine* 2007; **26**(27):4925–4938.
135. Zajicek JP, Hobart JC, Slade A, Barnes D, Mattison PG, MUSEC Research Group. Multiple sclerosis and extract of cannabis: results of the MUSEC trial. *Journal of Neurology Neurosurgery and Psychiatry* 2012; **83**(11):1125–1132.
136. Zeymer U, Suryapranata H, Monassier JP, Opolski G, Davies J, Rasmanis G, Linssen G, Tebbe U, Schröder R, Tiemann R, Machnig T, Neuhaus KL. The Na+/H+ exchange inhibitor eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction. Results of the evaluation of the safety and cardioprotective effects of eniporide in acute myocardial infarction (escami) trial. *Journal of the American College of Cardiology* 2001; **38**(6):E1644–E1650.
137. Morgan C, Huyck S, Jenkins M, Chen L, Bedding A, Coffey C, Gaydos B, Wathen J. Adaptive design: Results of 2012 survey on perception and use. *Therapeutic Innovation & Regulatory Science* 2014; **48**:473–481.
138. Donohue JF, Fogarty C, Lötvall J. Once-daily bronchodilators for chronic obstructive pulmonary disease: indacaterol versus tiotropium. *American Journal of Respiratory and Critical Care Medicine* 2010; **182**:155–162.
139. Barnes PJ, Pocock SJ, Magnussen H. Integrating indacaterol dose selection in a clinical study in copd using an adaptive seamless design. *Pulmonary Pharmacology & Therapeutics* 2010; **23**:165–171.
140. Gallo P, DeMets D, LaVange L. Considerations for interim analyses in adaptive trials, and perspectives on the use of DMCs. In *Practical Considerations for Adaptive Trial Design and Implementation*, He W, Pinheiro J, Kuznetsova OM (eds). Springer: New York, Heidelberg, Dordrecht, London, 2014; 259–272.

141. Lawrence D, Bretz F, Pocock S. INHANCE: An adaptive confirmatory study with dose selection at interim. In *Indacaterol - The First Once-Daily Long-Acting Beta2 Agonist for COPD*, Trifilieff A (ed.) Springer: Basel, 2014; 77–93.
142. Lawrence D, Bretz F. Approaches for optimal dose selection for adaptive design trials. In *Practical Considerations for Adaptive Trial Design and Implementation*, He W, Pinheiro J, Kuznetsova OM (eds). Springer: New York, Heidelberg, Dordrecht, London, 2014; 77–93.
143. Léauté-Labrèze C, Hoeger P, Mazereeuw-Hautier J, Guibaud L, Baselga E, Posiunas G, Phillips RJ, Caceres H, Lopez Gutierrez JC, Ballona R, Friedlander SF, Powell J, Perek D, Metz B, Barbarot S, Maruani A, Szalai ZZ, Krol A, Boccara O, Foelster-Holst R, Febrer Bosch MI, Su J, Buckova H, Torrelo A, Cambazard F, Grantzow R, Wargon O, Wyrzykowski D, Roessler J, Bernabeu-Wittel J, Valencia AM, Przewratil P, Glick S, Pope E, Birchall N, Benjamin L, Mancini AJ, Vabres P, Souteyrand P, Frieden IJ, Berul CI, Mehta CR, Prey S, Boralevi F, Morgan CC, Heritier S, Delarue A, Voisard JJ. A randomized controlled trial of oral propranolol in infantile hemangioma. *New England Journal of Medicine* 2015; **72**(8):735–746.
144. Heritier S, Lô SN, Morgan CC. An adaptive confirmatory trial with treatment selection: practical experiences and unbalanced randomization. *Statistics in Medicine* 2011; **30**:1541–1554.
145. Heritier S, Lo SN, Voisard JJ, Gautier S, Morgan-Bourniol C. A single pivotal adaptive trial in infants with proliferating hemangioma: rationale, challenges, experience and recommendations. In *Modern Adaptive Randomized Clinical Trials: Statistical, Operational, and Regulatory Aspects*. Chapman & Hall/CRC Press: Boca Raton, London, New York, 2015.
146. Chaturvedi PR, Antonijevic Z, Mehta C. Practical considerations for a two-stage confirmatory adaptive clinical trial design and its implementation: ADVENT trial. In *Practical Considerations for Adaptive Trial Design and Implementation*, He W, Pinheiro J, Kuznetsova OM (eds). Springer: New York, Heidelberg, Dordrecht, London, 2014; 77–93.
147. Hünseler C, Balling G, Röhlig C, Blickheuser R, Trieschmann U, Lieser U, Dohna-Schwake C, Gebauer C, Möller O, Hering F, T. Hoehn, Schubert S, Hentschel R, Huth R G, Müller A, Müller C, Wassmer G, Hahn M, Harnischmacher U, Behr J, Roth B. Continuous infusion of clonidine in ventilated newborns and infants: A randomized controlled trial. *Pediatric Critical Care Medicine* 2014; **15**:511–522.
148. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; **40**:1079–1087.
149. Lehmacher W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* 1991; **47**(2):511–521.
150. Wassmer G, Reitmeir P, Kieser M, Lehmacher W. Procedures for testing multiple endpoints in clinical trials: An overview. *Journal of Statistical Planning and Inference* 1999; **82**:69–81.
151. EMA. *Guideline on Clinical Trials in Small Populations (chmp/ewp/83561/2005)*. European Medicines Agency: London, UK, 2006. http://www.ema.europa.eu.
152. Emerson SS. Statistical packages for group sequential methods. *The American Statistician* 1996; **50**:183–192.
153. Wassmer G, Vandemeulebroecke M. A brief review on software developments for group sequential and adaptive designs. *Biometrical Journal* 2006; **48**(4):732–737.
154. Zhu L, Ni L, Yao B. Group sequential methods and software applications. *The American Statistician* 2011; **65**:127–135.
155. Tymofyeyev Y. A review of available software and capabilities for adaptive designs. In *Practical Considerations for Adaptive Trial Design And Implementation*, He W, Pinheiro J, Kuznetsova OM (eds). Springer: New York, Heidelberg, Dordrecht, London, 2014; 139–155.
156. Pahl R. *GroupSeq: A GUI-based program to compute probabilities regarding group sequential designs*, 2014. http://cran.rproject.org/web/packages/GroupSeq; R package version 1.3.2.
157. Anderson K. *gsDesign: Group sequential design*, 2014. http://cran.r-project.org/web/packages/gsDesign; R package version 2.9–3.
158. Casper C, Perez OA. *ldbounds: Lan-DeMets method for group sequential boundaries*, 2014. http://cran.r-project.org/web/packages/ldbounds; R package version 1.1–1.
159. Izmirlian G. *Pwrgsd: Power in a group sequential design*, 2014. http://cran.r-project.org/web/packages/PwrGSD; R package version 2.0.
160. Schoenfeld DA. *seqmon: Sequential monitoring of clinical trials*, 2012. http://cran.r-project.org/web/packages/seqmon; R package version 0.2.
161. Vandemeulebroecke M. *adapttest: Adaptive two-stage tests*, 2009. http://cran.r-project.org/web/packages/adapttest; R package version 1.0.
162. Hack N, Brannath W, Brückner M. AGSDest: Estimation in adaptive group sequential trials 2013. http://cran.r-project.org/web/packages/AGSDest; R package version 2.1.
163. Parsons N. asd: Simulations for adaptive seamless designs. *http:* 2013. http://cran.r-project.org/web/packages/asd; R package version 2.0.
164. Rohmeyer K, Klinglmüller F. gMCP: Graph based multiple comparison procedures 2014. http://cran.r-project.org/web/packages/gMCP; R package version 0.8–7.
165. Fisher A, Rosenblum M, Jaffee H. interAdapt 2014. http://cran.r-project.org/web/packages/interAdapt; R package version 0.1.
166. Chang M. *Adaptive Design Theory and Implementation Using SAS and R*. Chapman and Hall/CRC: Boca Raton, London, New York, 2008.
167. Brannath W, Koenig F, Bauer P. Multiplicity and flexibility in clinical trials. *Pharmaceutical Statistics* 2007; **6**(3): 205–216.
168. Burman C-F, Lisovskaja V. The dual test: Safeguarding p-value combination tests for adaptive designs. *Statistics in Medicine* 2010; **29**(7-8):797–807.
169. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine* 2011; **30**(28):3267–3284.
170. Glimm E. Comments on Adaptive increase in sample size when interim results are promising: A practical guide with examples by CR Mehta and SJ Pocock. *Statistics in Medicine* 2012; **31**(1):98–99.

171. Gould AL. Interim analysis for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine* 1992; **11**:53–66.
172. Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics - Theory and Methods* 1992; **21**:2833–2853.
173. Gould AL. Planning and revising the sample size for a trial. *Statistics in Medicine* 1995; **14**:1039–1051.
174. Kieser M, Friede T. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* 2000; **19**:901–911.
175. Bauer P, Brannath W. The advantages and disadvantages of adaptive designs for clinical trials. *Drug Discovery Today* 2004; **9**:351–357.
176. Quinlan JA, Krams M. Implementing adaptive designs: logistical and operational considerations. *Drug Information Journal* 2006; **40**:437–444.
177. Gaydos B, Anderson KM, Berry D, Burnham N, Chuang-Stein C, Dudinak J, Fardipour P, Gallo P, Givens S, Lewis R, Maca J, Pinheiro J, Pritchett Y, Krams M. Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Information Journal* 2009; **43**:539–565.
178. Antonijevic Z, Gallo G, Chuang-Stein C, Dragalin V, Loewy J, Menon S, Miller ER, Morgan CC, Sanchez M. Views on emerging issues pertaining to data monitoring committees for adaptive trials. *Therapeutic Innovation & Regulatory Science* 2013; **47**:495–502.
179. Bauer P, König F. The risks of methodology aversion in drug regulation. *Nature Reviews Drug Discovery* 2014; **13**(5): 317–318.

