Application of Adaptive Designs – a Review

P. Bauer * and **J. Einfalt**

Section of Medical Statistics, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

Received 28 November 2005, revised 12 December 2005, accepted 23 December 2005

Summary

A literature search has been performed to review applications of the adaptive design methodology based on the combination test or conditional error function approach. Some features of the 60 papers identified are summarized, e.g., the specific methodology used, calendar year, country, impact factor of the journal, number of planned and performed stages respectively, stopping for futility boundaries, type of adaptations and others. A selection of the ten recent publications in journals with the highest impact factors is discussed in more detail.

Most applications up to now aim at sample size reassessment, the majority of papers is coming from Germany. Although we found that renowned journals allow for sufficient space to present the new statistical methodology in all its necessary details, the general impression is that the presentation of the adaptive designs methodology in applied papers has to be improved. Education and development of standards could help to achieve this.

Key words: Adaptive designs; Combination test; Conditional error function; Applications; Literature review.

Supperting information for this article is available on the WWW under http://www.wiley-vch.de/contents/jc_2221/2006/200510204_s.pdf

1 Introduction

In the last 20 years several methods have been proposed how to perform mid-trial design modifications in multi-stage designs without compromising on the type I error rate. The basic motivation behind these proposals was to include learning from experience into the course of a trial, so that designs can be improved whenever evidence for misspecifications in the planning phase arises from inside or outside the running trial. There was also a certain disagreement with the common practice to deal with design modifications simply by writing amendments to the trial protocol. Instead in adaptive (or flexible) designs the option of a design modification is considered as an intrinsic quality which may improve on the performance of the trial.

Essentially there are two concepts behind these recent developments:

1. The combination test principle uses stage-wise test statistics which are combined according to a pre-defined combination function (Bauer, 1989; Bauer and Köhne, 1994)
2. The conditional error principle states, that any type of design modifications can be performed at any time of the trial as long as the conditional error of the new design does not exceed the conditional error of the pre-planned design (Proschan and Hunsberger, 1995; Müller and Schäfer, 2001, 2004).

* Corresponding author: e-mail: Peter.Bauer@meduniwien.ac.at, Phone: ++43-1-404007489

Both methods are closely related (Posch and Bauer, 1999) and allow flexibility with regard to, e.g., the number of interim looks, forthcoming decision boundaries, sample sizes, sample size allocation, dropping or adding of treatments, but even with regard to a modification of hypotheses. Adaptive designs can be planned in such a way, so that in case of completing the trial without performing any design modifications statistical analysis is identical to a conventional analysis in group sequential designs. The main reason for the critical discussion in the literature is that in case of performing design modifications the price to be paid for flexibility may be the use of non-standard test statistics other than the common "sufficient" test statistics.

Critical comments on the use of adaptive designs are made also from a regulatory perspective (Koch, 2006). Issues of concern are, e.g., the problems of maintaining the integrity and persuasiveness of results obtained after substantial changes and the complications arising in estimation following adaptive designs. The seemingly fundamental advantage of adaptive designs that the rules for design modifications need not to be laid down a priori (only assuming that the rules applied are measurable) also creates a problem: The sample space is not defined in advance so that concepts such as unbiased estimation can not be applied. There is no such thing as free lunch.

Here we do not further discuss methods which have been considered for dealing with problems like the application of non-standard test statistics, estimation and multiple inference in adaptive designs. Instead we try to investigate the impact of adaptive design methodology on the applied field in medicine. Is the methodology applied to a noticeable extent, what are the applications, and are the methods applied and presented properly?

2 The Literature Search

The notion "adaptive" has been used in the statistical literature before, e.g., for certain randomization procedures such as the play-the-winner rule. Since we were interested to look at the application of a specific type of methodology we selected a list of 75 papers dealing with the methodology of adaptive designs based on the combination test or conditional error function principle between 1989 and 2004. As a consequence we did not search for applications with design adaptations using a Bayesian type of inference. Most of the methodological papers we have been interested in have been published in statistical journals, however, some of them are rather general overviews published in medical journals. The latter have been included in the search list since they may have stimulated applications. The list may not cover all contributions ever published on adaptive designs based on the combination test or the conditional error principle. But to our judgement this list covers those papers which would have been cited if somebody had performed an adaptive design of the type of interest in the past. More recent literature will hardly lead to an application published before 31st of August 2005 (which was the deadline of the search in the applied literature. The list of 75 papers behind the search can be found in the supporting information.

The search for applied papers has been performed in the ISI Web of Science (Science Citation Index Expanded, Information Social Sciences Citation Index, Information Arts & Humanities Citation Index) asking for all published papers which contain a reference to at least one of the 75 "methodological" papers ("cited reference search"). Some of the methodological papers in statistical journals contain data at least to demonstrate how to perform the calculations for adaptive designs. Since we wanted to look at the impact of these publications on the applied field they were not included in the review. We are convinced that results of an adaptive study will not only be published as an accompanying example in a methodological paper, so that the main publication of the study would be identified by the search anyway. Moreover only papers were considered which contain real study data, excluding 4 papers presenting "planned" designs and 5 papers mentioning the adaptive methodology, e.g., in the discussion, but not applying it. Finally 60 applied papers were identified fulfilling the selection criteria. All these papers are from areas related to medicine. The complete list of papers can be found in the supporting information.

3 Variables Extracted from the Identified Papers

The 60 papers were analysed with regard to the following characteristic:

year
country (of the corresponding author)
category of the trial/journal
impact factor 2004
method
number of planned stages/interim analyses
number of stages/interim analyses carried out
test procedure for H_0 in stage 1
level of significance (level α) of the whole trial (including formulation: 1 or 2 sided)
power
stopping for futility boundary
early rejection boundary
rejection boundary for combination test at the end of the trial
planned adaptations
adaptations carried out
reason for early termination
result at the end of the trial
planned sample size of the first stage
actual sample size of the first stage
planned sample size of the second stage
actual sample size of the second stage
planned total sample size
actual total sample size
ITT sample size
PP sample size
sample size from which the results in the paper were calculated

Some descriptive statistics of these variables will be given in section 4.1.

Finally the 10 papers published between 2003 and 2005 in the journals with the highest impact factors were selected and a more detailed evaluation of the specific adaptive design methodology applied in these papers was undertaken. The narratives for this type of evaluation will be given in Section 4.2.

A discussion of the results will be given in the closing Section 5.

4 Results

4.1 Descriptive statistics

Figure 1 shows the number of published applications depending on the calendar year of publication including the method applied. There are two points to be considered: The number of applications shows an increasing tendency over the years, but the increase is much smaller than one would expect for a booming methodology (the year 2005 covers only 8 months up to the time of the literature search, and due to delayed input into the Web of Science a number of papers published in 2005 up to August will be missing). The most widely used methodology is based on the Fisher combination test for p-values, as proposed by Bauer and Köhne (1994) ($n = 48$), followed by the method using the inverse normal combination function of Lehmacher and Wassmer (1999) ($n = 5$) and the conditional error function approach by Proschan and Hunsberger (1995) ($n = 4$). One paper each referred to Fisher's combination test as applied by Bauer and Röhmel (1995) and one to Cui et al. (1999), which essentially is an application of the inverse normal combination test. In one paper there were two

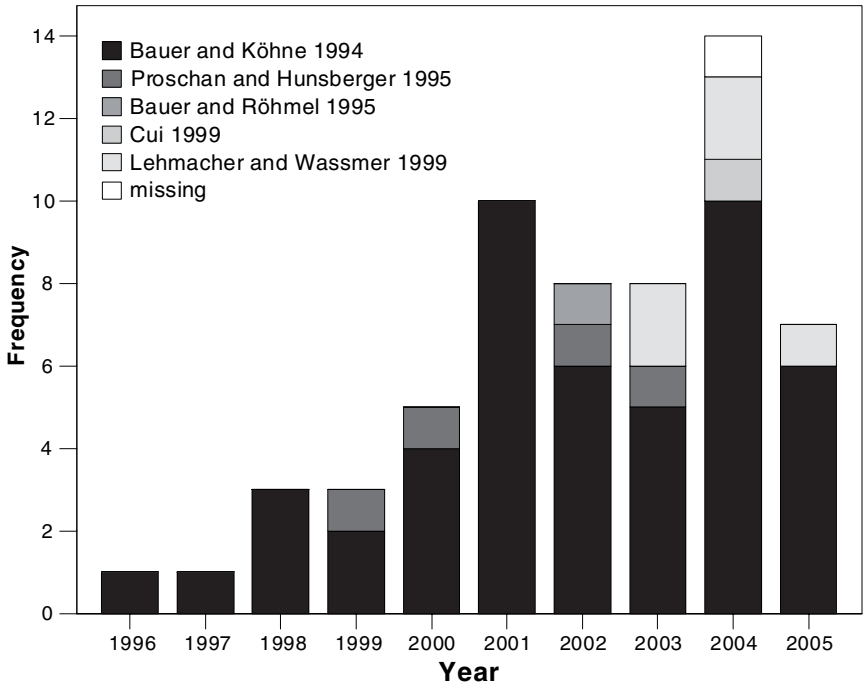

Figure 1 Number of published applications by calendar year and the adaptive methodology applied.

references to different approaches and it was not clear which method and how it had been used. Apparently in application there is a preference to use the product of p-values to combine information from different stages of a trial. This may be caused by the appealing simplicity of this criterion. In the original paper (Bauer, 1989) the arguments for the adaptive design methodology have been derived generally referring to general combination tests ("The proposed test then rejects the global null hypothesis ... if a (preassigned) level-α-test for combining k p-values from independent samples is rejected on the basis of $p_1 \ldots, p_k$. Also in the paper of Bauer and Köhne (1994) we find: "There are numerous strategies for testing the intersection H_0 of two (or k) individual null hypotheses based on independently and uniformly distributed p-values for the individual null hypotheses (...). Fisher's criterion using the product of p-values has good properties, is well known, and is used here as an example." Meanwhile it has been pointed out repeatedly that the normal combination function has favourable properties because of its close relationship to group sequential designs: Without performing any adaptations, the statistical analysis of the adaptive design based on the inverse normal function is equivalent to the conventional statistical analysis of a group sequential design (Müller and Schäfer, 2001). But still the applied community seems to follow the parsimonious method which has been used as an example in the initial papers.

A Box-plot of the impact factors (2004) ascribed to the journals containing the publications is given in Figure 2 showing a large variety. The 60 papers also come from many different medical fields. Figure 3 gives the distribution of publications depending on the country of the corresponding author. Here there is an overwhelming dominance of publications from Germany.

The number of planned and performed stages is shown in Table 1. Here only in one study the number of planned stages is missing, whereas the number of stages actually performed was missing in 4 studies. We found two studies with only a single planned stage (an example is discussed below). Three studies were planned for 3 stages, one for 4 and two for 5 stages. In none of these studies the number of stages performed, if mentioned in the paper, is larger than the number of stages planned

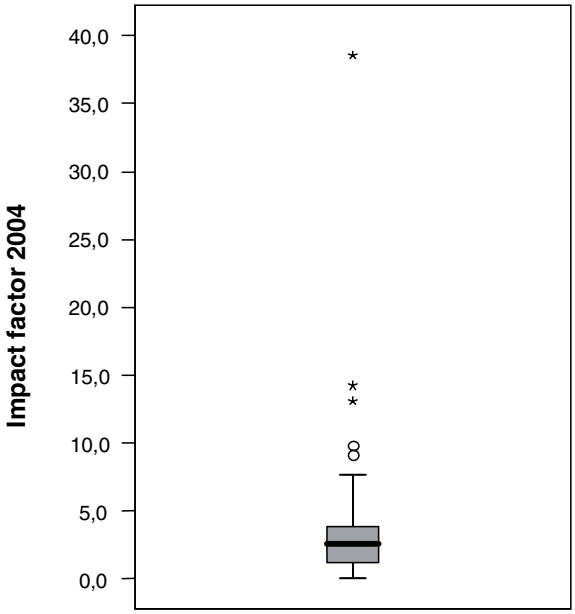

Figure 2 Impact factors (in the year 2004) of the medical journals which published applications of adaptive designs.

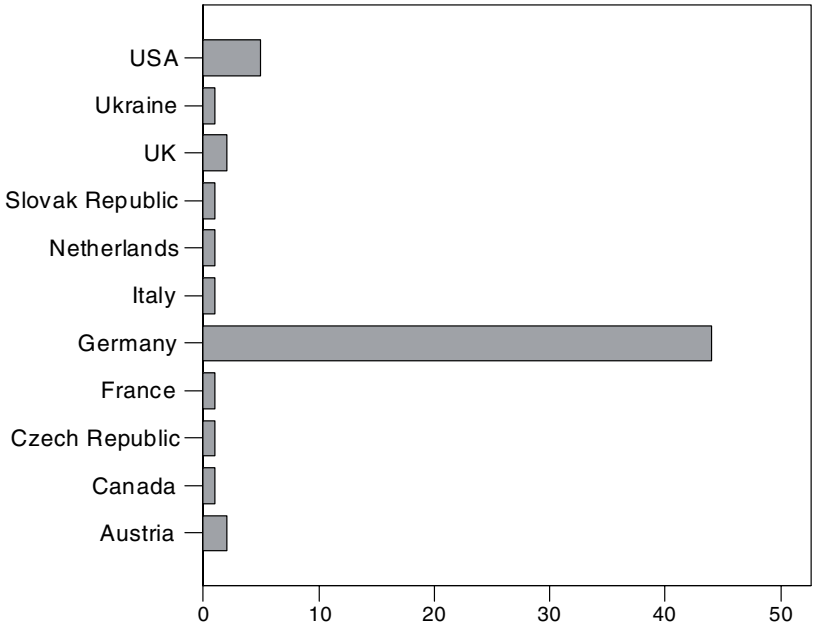

Figure 3 Number of publications depending on the country of the corresponding author.

Table 1 Number of stages planned and actually carried out.

Number of stages planned	Number of stages carried out				Total
	1	2	3	missing	
1	2	0	0	0	2
2	37	14	0	0	51
3	1	0	0	2	3
4	0	1	0	0	1
5	0	0	1	1	2
missing	0	0	0	1	1
Total	40	15	1	4	60

(although this is an option of adaptive designs: Müller and Schäfer, 2004; Brannath et al., 2002). However, in a single study planned for two stages where an additional second interim analysis has been introduced the third stage has not been performed.

The overall significance level α chosen for the designs are shown in Table 2. In three cases this quantity was not mentioned in the paper. In the majority of cases (34) it could not be recovered whether a one-sided or two-sided test formulation has been chosen. In the three cases applying a significance level of 0.025 without stating if one- or two-sided it is rather likely that a one-sided test was intended. In the 31 cases applying a significance level of 0.05 one generally would assume that by convention a two-sided test has been applied. However, the adaptive methodology originally was mainly formulated in terms of one-sided tests to avoid that conflicting directional effects at the different stages combine to a single final directional test decision. Therefore it is questionable if only two-sided tests have been planned in these 31 cases. This lack of information should not be specific for the application of flexible designs, since incomplete descriptions of the statistical methodology will be also encountered in other types of statistical methods applied in medicine.

In Table 3 the stopping for futility boundaries α_0 planned for the first stage p-values are shown. In 12 cases there was no information if an early stopping for futility has been pre-planned. In 14 cases it was clear that no stopping for futility was pre-planned ($\alpha_0 = 1$). This means that in the majority of studies (34) a stopping for futility option has been pre-planned. This is certainly an unusually high proportion. It may be explained by the importance which has been ascribed to early stopping for futility in the initial papers. There, stopping for futility was suggested whenever stage-wise effects are far away from the expected trend. The mostly used boundary ($\alpha_0 = 0.5$) in the one-sided normal case means that the trial is stopped after the first stage if an effect in the wrong direction has been observed. Such a choice looks appealing and has been suggested also for other sequential procedures. In general, however, using mandatory stopping for futility boundaries to achieve narrower decision

Table 2 Overall significance level α of the entire trial.

overall level α	1-/2-sided	n
0.025	1-sided	10
0.025	missing	3
0.05	1-sided	9
0.05	2-sided	3
0.05	missing	31
0.052	2-sided	1
missing	missing	3

Table 3 Stopping for futility boundaries α_0 for the first stage *p*-value.

Stopping for futility boundary	*n*
0.2	3
0.3	7
0.4	3
0.5	14
0.6	6
0.7	1
1	14
missing	12

boundaries in the final analysis does not lead to a sufficient gain in power for the price to be paid in terms of flexibility. If stopping for futility boundaries are as small as $\alpha_0 = 0.2$ (Table 3), this may lead to a loss of power in case of small first stage sample sizes, or in case of a-priori misspecifications, e.g., when underestimating the variance in the planning phase (Posch and Bauer, 2001)

The critical boundaries α_1 for the first stage *p*-values to achieve early rejection of the null hypothesis after the first stage (as a fraction of the overall level α) are given in form of a box-plot in Figure 4. In 16 cases this fraction could not be recovered from the paper. Half of the fractions are between 0.4 and 0.466, minimum and maximum are 0.104 and 0.696 respectively: This variation is not surprising for sequential designs having in mind that, e.g., there is a huge difference in early stopping boundaries between Pocock and O'Brien-Fleming type boundaries. The large local first stage

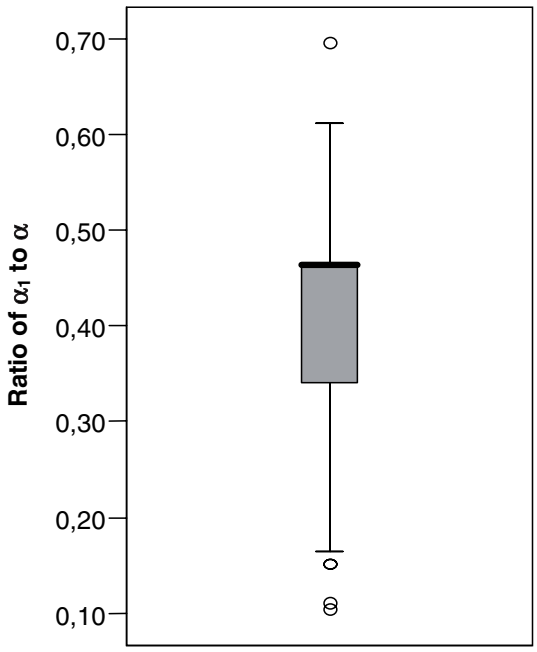

Figure 4 Ratio of the critical boundaries α_1 for the first stage p-values to the respective level α of the each trial ($n = 44$).

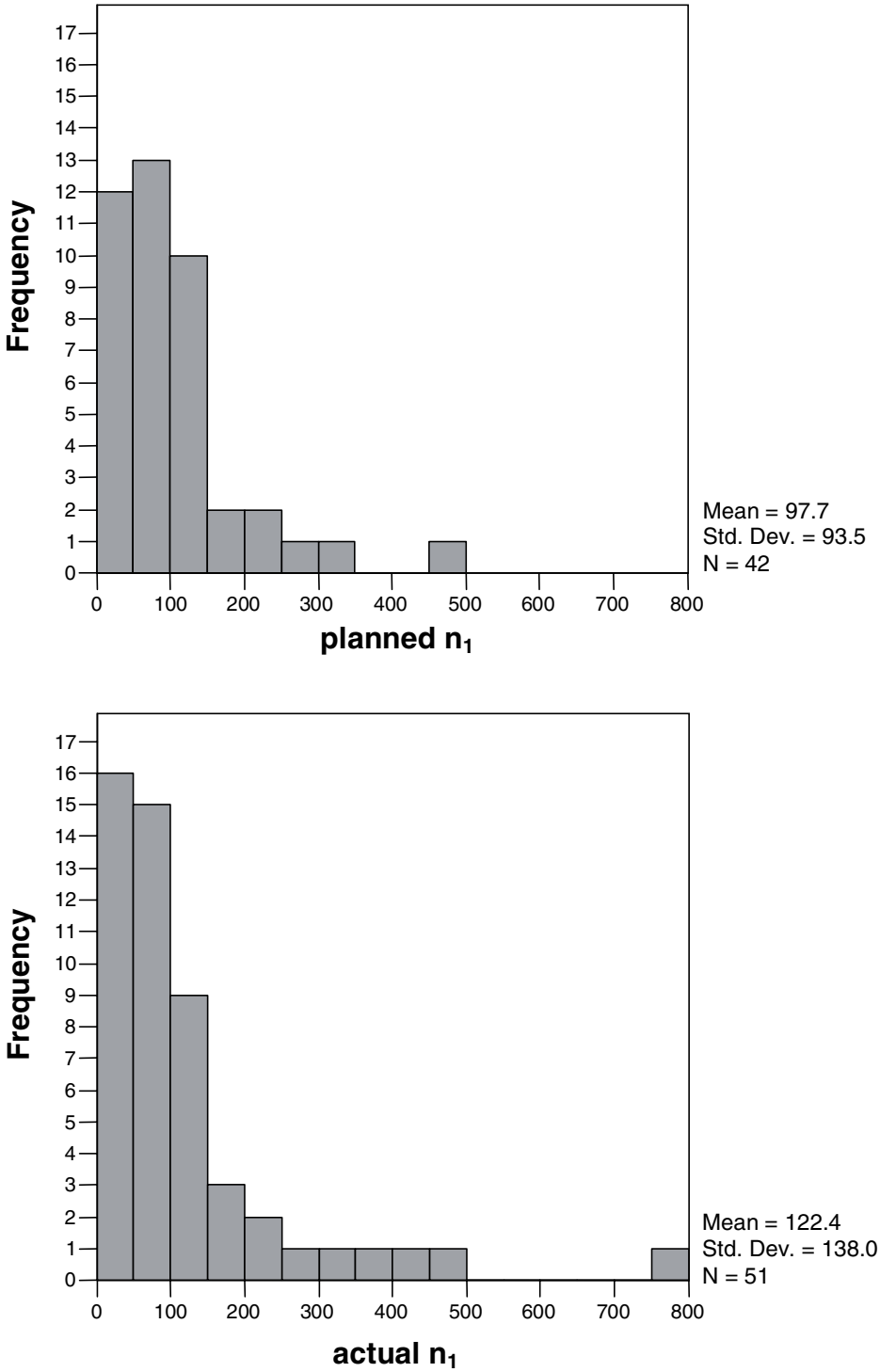

Figure 5 Histograms of planned and actually performed sample sizes at the first stage (absolute frequencies).

rejection boundaries α_1 are obviously related to the small mandatory stopping for futility boundaries α_0 in Table 3.

As a first stage test mostly applied has been the rank-sum U-test (16 studies), followed by Fisher's exact test (9) and the χ^2-test (7). One might be surprised that the t-test is not on this leading list for adaptive studies. It may be that in case of a normal distribution experimenter hesitate to apply methods beyond the "engraved" standard methods.

The a-priori planned total sample size could be recovered in eighteen studies (median 200, ranging from 50 to 1150). But this type of information is also frequently missing in other types of designs. Figure 5 shows the distribution of the planned (from 42 studies) and actually performed sample sizes (from 51 studies) at the first stage. It is not unexpected that the planned sample size of the first stage is more often missing than the actually performed sample size. There are 13 studies where only the planned sample size is missing, 4 where only the sample size actually applied is missing and 5 studies where both are missing. For the 38 studies where both quantities have been given the distribution of the difference between the actually performed and the planned first stage sample size is fairly symmetric around zero (median 0, mean -2.1, minmum -70, maximum 43) indicating that there is no systematic tendency to apply first stage sample sizes smaller than planned.

Altogether 40 studies ended after the first stage, 12 stopping for futility, the rest leading to early rejections. In two studies it was not clear if the study was continued. Among the remaining 18 studies proceeding to the second stage there were 5 with missing information on the type of adaptations performed, in 10 of the studies only sample size adaptations have been performed, one study each applied sample size reassessment plus dropping a treatment arm or adjusting a dose respectively. In one trial the adaptation covered sample size reassessment and insertion of a further interim analysis.

4.2 The ten papers in the highest ranked journals

In the following discussion of the ten top ranked papers published between 2003 and 2005 we are aware that presenting statistical methods in medical journals is a big challenge. On the one hand there is a heavy pressure on space, on the other hand there is a need to explain the methodology, particularly if it is innovative. Here we were interested how this issue is handled in the new area of adaptive designs.

1. A two-armed, randomized, double blind, placebo-controlled multi-centre clinical trial (Taylor et al., 2004) has been performed to examine whether a therapy provides benefit in a subgroup previously noted to have a favourable response to this therapy. The reason to apply an adaptive design was explained: "Since the composite measure used in this trial had not been evaluated in previous trials, two interim analyses were prespecified in the protocol to permit an assessment of the adequacy of the sample size without knowledge of efficacy." The pre-planned method was that of Cui et al. (1999). "The initial estimate that 800 patients (400 per group) were needed for the study to have sufficient statistical power ($P < 0.02$) was modified to 1100 patients (550 per group) ... on the basis of the prespecified interim analysis." "The decision to stop the trial was based on the Lan-DeMets sequential boundaries. The trial was halted owing to a significantly higher mortality rate in the placebo group ... ." The Kaplan-Meier survival analysis apparently has been done in the classical way not accounting for the adaptive interim analysis.
2. A two-armed, randomized, double blind, placebo-controlled multi-centre clinical trial (Isenmann et al., 2004) has been performed to investigate if a prophylactic therapy reduces the proportion of patients with a certain type of event. A sample size of 100 patients in each group was precalculated (at the one-sided level 0.05). "An adaptive interim analysis (according to Bauer and Koehne with $\alpha_0 = 0.5$) was performed for the primary endpoint ... after 105 patients had been enrolled. A χ^2–test ($p = 0.719$, one-sided) was performed: "Consequently, recruitment was stopped because the trend in the incidences was in the opposite direction and the final analysis of the study data was performed". Hence, this study ended with a stopping for futility decision.

3. A two-armed, randomized, multi-centre trial (Stahl et al., 2005) has been performed to investigate if a therapy is non-inferior to this therapy in combination with an additional one. Sample size planning was based on a one-sided level of 0.05 to show "equivalence between treatment groups assessed by the one-sided log-rank test according to Wellek, with a minimum acceptable difference ... of $\delta = -0.15$. ... From the literature and own data we estimated a 2-year survival rate of approximately 35% in arm A ... resulting in 2×100 patients necessary to accept or decline equivalence" (power 80%, non-inferiority margin 15%). "The interim analysis of the first 119 eligible, randomized patients in the adaptive scheme (Bauer & Koehne, 1994) showed that, for the log-rank test for equivalence of overall survival, a total of 175 patients had to be allocated." After this reduction of the pre-planned sample size (in an early interim analysis) a conventional statistical analysis has been performed for the patients recruited in the final evaluation: "Overall survival at two years was equivalent between both treatment groups ... (log-rank test for equivalence with $\delta = -0.15$, $P = 0.007$)." Without going into any details of the analysis the crucial point in this study was that there was a long follow up period going beyond the two years (median follow up time 6 years). Overall survival in this study diverted substantially between the treatment arms after 2 years which was supported by a clear superiority of the combination therapy in a secondary endpoint. So the conclusion in the abstract "... showed overall survival to be equivalent between the two treatment groups" is only based on a part of the collected data. This problem is not due to the application of adaptive designs.
4. This study (Szegedi et al., 2005) reports the results of a two-armed, randomized, double blind multi-centre clinical trial to compare the efficacy of two active drug treatments. It has been planned as an adaptive two-stage design according to Fisher's combination test. Nested hypotheses have been considered to have the option of switching from non-inferiority to superiority of the test treatment as compared the active control. All necessary details of the design parameters are given: An a-priori sample size calculation for the first stage to achieve a certain probability for a trend in favour of non-inferiority already in the interim analysis ($n_1 = 50$ patients in each treatment group), boundaries for early rejection ($\alpha_1 = 0.01$) and stopping for futility ($\alpha_0 = 0.5$) are given as well as the way of reassessing the sample size in the interim results (total sample size 75 patients in each group). A one-sided confidence interval for the effect measure is given accounting for the interim look and a conventional statistical analysis is added leading to the same conclusion of superiority of the test treatment. This paper is a very good example of how application and presentation of adaptive designs could be done appropriately. It also demonstrates that high rank journals give the necessary extra space for detailed description of new statistical methods.
5. In this two-armed randomized, placebo controlled multi-centre clinical trial (Schaefer et al., 2004) it has been investigated if a new treatment is able to reduce a proportion of events. "The trial was planned according to the 2-stage Bauer and Köhne adaptive method, which combines the 2 p-values of the separate stages with Fisher's combination test. The first interim analysis was conducted after 58 patients had been treated per protocol. The sample size has been recalculated on the basis of the interim results. Furthermore, the recursive testing principle was introduced (Brannath et al., 2002), allowing an additional interim analysis. The second interim analysis was planned after a total of 130 patients." This paper exploits the feature of inserting additional interim analyses based on the mid-trial results. However, the third stage of the trial after the second interim analysis has not been performed: "At the time of the second interim analysis data from 155 randomized patients were available. Enrolment was suspended because the p-value for the difference in infection rates after 10 days exceeded the a priori threshold for stopping the trial for futility." A conventional statistical analysis was performed after terminating the trial.
6. In this two-armed, randomized, double blind and placebo-controlled clinical trial (Sperber et al., 2004) efficacy of a therapy to prevent a disease has been studied. "The study was designed as the first stage of a 2-stage adaptive design based on a methodology described by Bauer and

Köhne. Two co-primary endpoints were considered. When ∼50 subjects had completed the study and both end points were known for each subject, an adaptive interim statistical analysis of the data was performed to determine a final sample size for the study and to redefine the primary efficacy parameter. Early rejection based on non-stochastic curtailing (overall one-sided level of the product test, $\alpha = 0.05$) was pre-planned. In the interim analysis, one endpoint showed poor results. Due to the low first stage sample size (24 and 22 for verum and placebo respectively) there were very large standard confidence intervals. The trial was stopped for futility: "In view of the primary importance assigned by the study sponsor" to one of the two endpoints "the sponsor decided to terminate the study at stage 1." In the abstract the authors write: "Administration of ... did not decrease the rate of infection; however, because the small sample size, statistical hypothesis testing had a relatively poor power to detect statistically significant differences ...". It can not be judged here if the option of adapting the endpoint has been justifiable in this application (it has not been used anyway) or stopping for futility was reasonable. In any case, this does not seem to be the way how stopping for futility should be interpreted in practice.

7. In this two-armed, randomized multi-centre clinical trial (Reinhart et al., 2004) the application of a new treatment has been compared with a standard therapy. "The study was conducted according to the flexible adaptive design approach (Bauer and Köhne, 1994; Proschan and Hunsberger, 1995). An interim analysis was performed after 50 subjects passed the fourth treatment day to obtain a reliable estimate of the primary variable (responder rate) and to ensure that the study had enough power to detect predefined differences. Due to the preliminary nature of the study, allowance was also made for modifying the treatment population if evidence of clearly superior effect emerged in any of the predefined subgroups". In the interim analysis actually performed after 93 patients they found an effect lower than projected and switched recruitment to a subgroup with a more promising interim effect estimate. "After the interim analysis, both treatment groups were continued, but further patient enrolment was restricted to patients suffering from ...". A final conventional analysis after 143 patients failed to demonstrate a difference between the treatment arms. Here a rather uncritical use of flexibility has been made, the analysis not accounting for the adaptive design.
8. In this two-armed, randomized multi-centre study (Franz et al., 2004) the question has been addressed, if a certain diagnostic strategy to decide on the necessity of a prophylactic therapy reduces the proportion of therapy applications as compared to the previous standard procedure applied in the contributing centres. At the same time it has to be assured that the new procedure is non-inferior with regard to the proportion of initially missed diseases among the patients (absolute non-inferiority margin 3% for the difference between the proportions). A number of 1150 patients had been pre-calculated as the necessary total sample size. "A pre-defined adaptive interim analysis (with $\alpha_0 = 0.3$, reference to Bauer and Köhne) was performed after the first year of the trial when 356 patients had been enrolled". The proportion of patients receiving prophylaxis was clearly lower with the new diagnostic algorithm, $p < 0.0001$. ("The first study hypothesis was confirmed at this point"). Non-inferiority at the one-sided level of 0.05 could not be established at that time, although the trend was in favour of the new decision algorithm. "On the basis of these results, a new sample size calculation revealed that another 353 patients were required in each group to prove 1-sided equivalence ... ". A conventional one-sided confidence interval was calculated for the difference in the proportion of initially missed diseases after altogether 1291 patients but again failed to meet the a priori-defined margin to establish non-inferiority. In principle the application of sample size reassessment in this situation with limited knowledge on the proportions of initially missed diseases seems to be very reasonable (the observed percentages of 14.5% under the new procedure versus 17.3% in the standard procedure were markedly different to the a-priori assumed proportions of 4% versus 9%). There was already a hint on this discrepancy in the interim analysis (18% versus 27%). Here keeping the original non-inferiority margin was a hard goal to be met with proportions much higher than expected in the planning phase. It is a delicate issue to change the non-inferiority margin during

the trial. The authors kept to their earlier assumptions and in the conclusions they cautiously stated that "This diagnostic strategy seemed to be safe".

9. This was a register study (Dunlap et al., 2003) analysing data of 853 patients from a database. After that the conditional power was calculated to achieve significance for the main study goal if the sample size would be as large as 2700 (referring to the paper of. Proschan and Hunsberger, 1995). Hence this study was not really following an adaptive design.
10. In this two-armed, randomized, double blind and placebo-controlled multi-centre trial (Corrigan et al., 2005), Fisher's combination test was projected to be used for combining the evidence after the first year of the trial with the evidence after the second year. The one year interval is given by the seasonal nature of the disease. "The difference between active and placebo treatment ... increased from 26.6% after the first year to 48.4% after the second treatment period. Both differences were statistically significant with $P_1 = 0.0258$ for 2002 and $P_2 = 0.0177$ for 2003". It seems that the analysis was performed two times for the same total number of 144 patients. It is not intended to question the validity of the conclusions derived from the study. But the method was planned for p-values dependent under the null hypothesis, which is violating a basic assumption of adaptive designs.

5 Conclusions and Discussion

Taking the large number of publications in the medical literature, the first conclusion is that adaptive designs of the type considered in this paper are not widely used in practice. However, this is not surprising remembering the long time, e.g., group sequential methods have needed to become an accepted and widely applied tool in medical studies. The method is used mainly in Germany, which is due to the broad early research on this subject performed in this area. Adaptations in practice are rather limited to sample size reassessment. All the sophistications published in this area (dropping or adding treatment arms, skipping or inserting interim analyses, modifying endpoints, modifying test statistics, focussing on subgroups, ...) have not really entered the medical practice. This concentration on sample size reassessment is not surprising, since this is an old issue under continuous discussion. It is also understandable that experimenters are rather hesitant to address the other types of more radical adaptations, which have been ruled out in other methods up to now (but may be the ones which make adaptive designs really useful). Moreover, applying these adaptations appropriately to our opinion needs careful consideration of a multiplicity of problems which generally become trickier than in conventional designs.

Overall, the standard of presenting the statistical methods behind the adaptive designs in applied publications is low. We are aware that reviewers often force the authors to rigorously cut down the statistical methods section. If a conventional t-test in a single stage design is applied this does not need a lengthy description (which also may be a practical argument to use such a simple design). But with adaptive methods design parameters have to be known to understand the decision procedure. The topics addressed could be: A motivation why adaptive designs have been chosen for the particular medical problem to be studied (e.g., dropping inefficient and/or unsafe treatment arms to protect patients); the types of adaptations planned (e.g., sample size reallocation after dropping treatments); description of the initially planned trial, in particular of the first stage (test statistics and method of combining stages, stopping boundaries, stage-wise sample sizes); performed adaptations together with the motivation (e.g., inserting or skipping interim analyses, the latter without wasting type I error probability); test decisions with test statistics (also stage-wise?); estimates and CI accounting for the adaptive design; if possible also conventional estimates.

In the discussion of the ten top ranked papers we have seen that high rank journals with a good statistical reviewing procedure do give space to a careful explanation of innovative designs. This is the good news for users of such designs. The bad news is that exploiting the potential of adaptive designs in real life trials will generally increase the problems in presenting the framework behind, e.g., when

typically a multiplicity of testing issues has to be addressed. Quite obviously, in the published literature a problem exists with estimation in adaptive designs. In general we find conventional statistical analyses at the end of the trial. Also in group sequential design applications, often conventional estimates are given. However, in adaptive designs often test statistics diverging from the conventional test statistics may have to be used for the test decisions. Here additional research is required on the properties of suitable estimates following design adaptations and decisions, since for future applications of adaptive designs this will be a crucial point. Altogether properly applying flexible designs requires more input of statisticians throughout the studies, and creates an additional challenge of an appropriate presentation of the applied methods and the decision modalities.

However, we may need such designs: Ethical arguments and economical considerations may ask to react to interim evidence. Common sense tells us not to stick to plans laid down at a time with intrinsic shortage of knowledge on the specific question under study. We have to react if we are confronted with the unexpected. We may be correctly guided by the interim information, but we may also be misguided and have to pay the price of using unconventional test statistics with unfavourable properties. So it has to be questioned if early estimates of the effect size will generally guide us to the right track (e.g., Bauer and König, 2006). But here and now halfway through a trial, what should guide us better than current interim data form the trial itself?

Mid-trial design modification may have a negative impact on the persuasiveness and perception of the results. Therefore, flexible designs require a high degree of transparency of their decision procedures and logistics. One way to achieve improvements is education. Another important aid would be to develop standards for the different steps of planning, performing, adapting, analysing and presenting such designs.

References

Bauer, P. (1989). Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie* **20**, 130–148.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.

Bauer, P. and König, F. (2006). The reassessment of trial perspectives from interim data – a critical view. *Statistics in Medicine* **25** (1), 23–26.

Bauer, P. and Röhmel, J. (1995). An adaptive method for establishing a dose response relationship. *Statistics in Medicine* **14**, 1595–1607.

Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97**, 1–9.

Corrigan, C. J., Kettner, J., Doemer, C., Cromwell, O., and Narkus, A. (2005). Efficacy and safety of preseasonal-specific immunotherapy with an aluminium-adsorbed six-grass pollen allergoid. *Allergy* **60** (6), 801–807.

Cui, L., Hung, H. M. J., and Wang, S. (1999). Modificaton of sample size in group sequential clinical trials. *Biometrics* **55**, 321–324.

Dunlap, S. H., Mallemala, S., Sueta, C. A., Schwartz, T. A., and Adams, K. F. (2003). Survival rates are similar between African American and white patients with heart failure. *American Heart Journal* **146** (2), 265–272.

Franz, A. R., Bauer, K., Schalk, A., Garland, S. M., Bowman, E. D., Rex, K., Nyholm, C., Norman, M., Bougatef, A., Kron, M., Mihatsch, W. A., and Pohlandt, F. (Group Author(s): Int IL-8 Study Grp). (2004). Measurement of interleukin 8 in combination with C-reactive protein reduced unnecessary antibiotic therapy in newborn infants: a multicenter, randomized, controlled trial. *Pediatrics* **114** (1), 1–8.

Isenmann, R., Runzi, M., Kron, M., Kahl, S., Kraus, D., Jung, N., Maier, L., Malfertheiner, P., Goebell, H., and Beger, H. G. (Group Author(s): ASAP Study Grp). (2004). Prophylactic antibiotic treatment in patients with predicted severe acute pancreatitis: A placebo-controlled, double-blind trial. *Gastroenterology* **126** (4), 997–1004.

Koch, A. (2006). A regulatory view on adaptive designs in Phase III clinical trials. *Biometrical Journal*.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286–1290.

Müller, H. H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57**, 886–891.

Müller, H. H. and Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* **23** (16), 2497–2508.

Posch, M. and Bauer, P. (1999). Adaptive two stage design and the conditional error function. *Biometric Journal* **41**, 689–696.

Posch, M. and Bauer, P. (2001). Interim analysis and sample size reassessment. *Biometrics* **56**, 1170–1176.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–1324.

Reinhart, K., Meier-Hellmann, A., Beale, R., Forst, H., Boehm, D., Willatts, S., Rothe, K. F., Adolph, M., Hoffmann, J. E., Boehme, M., and Bredle, D. L. (Group Author(s): EASy-Study Grp). (2004). Open randomized phase II trial of an extracorporeal endotoxin adsorber in suspected gram-negative sepsis. *Critical Care Medicine* **32** (8), 1662–1668.

Schaefer, H., Engert, A., Grass, G., Mansmann, G., Wassmer, G., Hubel, K., Loehlein, D., Ulrich, B. C., Lippert, H., Knoefel, W. T., and Hoelscher, A. H. (2004). Perioperative granulocyte colony-stimulating factor does not prevent severe infections in patients undergoing esophagectomy for esophageal cancer – A randomized placebo-controlled clinical trial. *Annals Of Surgery* **240** (1), 68–75.

Sperber, S. J., Shah, L. P., Gilbert, R. D., Ritchey, T. W., and Monto, A. S. (2004). Echinacea purpurea for prevention of experimental rhinovirus colds. *Clinical Infectious Diseases* **38** (10), 1367–1371.

Stahl, M., Stuschke, M., Lehmann, N., Meyer, H. J., Walz, M. K., Seeber, S., Klump, B., Budach, W., Teichmann, R., Schmitt, M., Schmitt, G., Franke, C., and Wilke, H. (2005). Chemoradiation with and without surgery in patients with locally advanced squamous cell carcinoma of the esophagus. *Journal Of Clinical Oncology* **23** (10), 2310–2317.

Szegedi, A., Kohnen, R., Dienel, A., and Kieser, M. (2005). Acute treatment of moderate to severe depression with hypericum extract WS 5570 (St John's wort): randomised controlled double blind non-inferiority trial versus paroxetine. *British Medical Journal* **330** (7490), 503–506.

Taylor, A. L., Ziesche, S., Yancy, C., Carson, P., D'Agostino, R., Ferdinand, K., Taylor, M., Adams, K., Sabolinski, M., Worcel, M., and Cohn, J. N. (Group Author(s): African-Amer Heart Failure Trial I). (2004). Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *New England Journal Of Medicine* **351** (20), 2049–2057.

