Discussion

Methodological Developments vs. Regulatory Requirements

Peter Bauer*

Core Unit for Medical Statistics and Informatics, Section of Medical Statistics,
Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

Summary

This is a discussion of the following three papers appearing in this special issue on adaptive designs: 'FDA's critical path initiative: A perspective on contributions of biostatistics' by Robert T. O'Neill, 'A regulatory view on adaptive/flexible clinical trial design' by H. M. James Hung, Robert T. O'Neill, Sue-Jane Wang and John Lawrence; and 'Confirmatory clinical trials with an adaptive design' by Armin Koch.

Key words: Adaptive designs; Adaptive information design; Adaptive study designs; Biostatistics; Clinical trial simulation; Dropping a treatment arm; Drug safety; Estimation; Experimental design; FDA critical path; Guidance development; Multiple endpoints; Non-inferiority trials; Phase III clinical trials; Quantitative risk/safety assessment; Regulatory biostatistics; Regulatory evaluation; Sample size reassessment; Statistical efficiency; Statistical information; Superiority.

There are three papers in this special issue which consider adaptive designs from a regulatory perspective. O'Neill (2006) reports on the Critical Path Initiative launched by the FDA in 2005 "to improve upon the science infrastructure behind product development and evaluation". He states that "Guidance development and promulgation has long been a mechanism to bring predictability to the drug development and review process. But guidance alone is likely not sufficient. We may also need to revisit the effectiveness of the current process of how planning of individual clinical trials occurs, and evaluate how effective the current planning process for use of the accumulation of information during product development derived from the collection and/or the sequence of clinical trials". This refers to a serious problem. On the one hand regulatory offices are strongly asked to create close guidance for the development processes. On the other hand in an area where different interests are focusing close guidance can be an obstacle for the use of innovative methods. It is understandable that people in the pharmaceutical industry tend to adhere closely to regulatory guidance in order to avoid later blame for having promoted innovative methodology (if no success has been achieved at the end).

It is important that regulators themselves take the initiative to provoke a search for possible improvements to established procedures. Asking what contribution the field of biostatistics can make to improving the efficiency of the development and evaluation process, the author emphasizes the field of drug safety and quantitative risk assessment. In my own experience, it would already be a big step if the currently available statistical methodology were more consistently used in this area too. When trying to get some information on the correlation of efficacy and safety variables recently, I was surprised that little information is available in the literature (although plenty of data should exist from clinical trials). One overly simplistic explanation for this is that in drug companies analyses of efficacy and safety are usually performed by different groups of people.

* e-mail: peter.bauer@meduniwien.ac.at

The author points out specific issues with strong statistical content which have to be considered for improving the development processes: Active-control non-inferiority trials, multiple endpoints, missing data, adaptive designs and simulation of clinical trials. Identifying adaptive designs as a topic of current interest also from a regulatory perspective links this paper to the others in this special issue. It is not surprising that one of the goals of the initiative is to develop guidance on how the clinical trial community should deal with these issues (recognizing that some EMEA guidance documents on these topics already exist).

Two further papers (Hung et al., 2006; Koch, 2006) mainly deal with regulatory concerns about adaptive designs. The authors of the first paper from the FDA have contributed innovative methodological research within this area themselves and so their arguments deserve special attention. They report about the type of design adaptations FDA reviewers have encountered in regulatory applications during the last 5 years: Sample size reassessment, termination of treatment arm, change of the primary endpoint, change of the statistical test, change of the study objective such as from superiority to non-inferiority or vice versa and selection of subgroups based upon externally available studies. Most applications were discouraged at the protocol stage for various reasons discussed by the authors.

For sample size planning Hung et al. advocate the a-priori definition of a minimum clinically important effect (MCIE). They question whether there is sufficient justification for using designs including mid-trial sample size reassessment. One reason is that conditional power evaluated during the trial can be too unreliable. Another reason is that in the case of complete pre-specification of the adaptation criterion a more efficient fixed maximum information design may be available (which of course is true). However, we should not forget that fixing an MCIE is not an easy task out in the field and it may even change in an evolving medical environment. Overpowering group sequential studies by choosing relatively small a priori effect sizes (counting at early stopping in case of a larger effect) is another option which would lead us to a scenario of trials with generally large (and potentially prohibitive) maximum sample sizes. For such a planning philosophy the stopping for futility decision will have to be considered very carefully not to end up with very large average sample sizes under the null hypothesis. There may be situations in which it could be useful to increase the sample size (in at least one treatment group). Due to an unexpected safety issue arising throughout the trial, a larger sample size may be needed to perform an appropriate quantification of the risk-benefit relationship. The authors have concerns that "the sample size re-estimation does not depend on any future data". This would be connected with breaking of the rules and like any other misconduct it has to be excluded by a careful documentation of the adaptation. Basically, I agree with the authors that mid-trial sample size reassessment should not be proposed as the standard in clinical trials although most published applications in medicine apply this type of adaptation (Bauer and Einfalt, 2006).

Dropping treatment/dose arms because of lack of efficacy and/or toxicity is an important application of adaptive designs. The authors concede that it may be advisable to redistribute the remaining planned sample size of a terminated arm to the remaining treatment arms (coupled with a proper adaptive test). It has been shown for genetic studies with a large number of hypotheses, that dropping the majority of hypotheses in an adaptive interim analysis and redistributing the remaining resources to the few selected hypotheses, may massively increase the power as compared to conventional designs (Zehetmayer, Bauer and Posch, 2005); see also the results in König, Bauer and Brannath (2006) for clinical trials.

For the situation of switching between superiority and non-inferiority the authors ask for pre-specification of a non-inferiority margin which should not depend on the outcome. However, adaptive designs would also permit an unforeseen mid-trial upgrade of the design objective, in the sequel going for superiority to the active control by more than a certain margin. For the methodology behind the construction of the underlying confidence intervals see Bauer and Kieser (1996).

The authors concede that a trial could fail because the primary endpoint is not a good choice. Understandably, they see little reason to advocate as serious an adaptation as a change of the primary endpoint in confirmatory trials. However, I do not believe that their comment on the good performance of a Bonferroni-adjusted multiple test procedure as compared to the adaptive test after switching between two endpoints can be generalized.

The problems of using non-standard test statistics and how to get appropriate estimates are critical issues. Here the many options offered by the design have to be paid for because without pre-specification of the adaptation rule no sample space is defined. There have been solutions proposed, e.g., by using dual tests which also meet the rejection criterion of a conventional test. We also already know a few things about the bias of conventional estimates when sample size reassessment is applied, but further research is necessary (see e.g., Brannath, König and Bauer, 2006).

The authors give some case studies as examples of what they call analysis problems arising in adaptive designs. Case study #1 refers to a clinical trial with two endpoints (surrogate and primary) which are analysed within the framework of a group sequential design. Statistically, this seems to be a nasty problem of multiple testing not specific to adaptive designs. The main problem in Case study #2 refers to a multiple test problem with a composite endpoint in adaptive designs, which can be treated by the closed testing principle (e.g., Kieser and Bauer 1999). Case study #3 is again a sequential multiple test problem including a surrogate endpoint of morbidity with a primary endpoint of mortality. Case study #4 refers to the combination of evidence from different phases of drug development where many changes are likely to occur. Since its general vague formulation covers a wide field of adaptations over various stages, inference on individual endpoints adjusting for multiplicity, if at all possible, will get increasingly complicated and at the same time increasingly less convincing.

The authors finally point out the many logistic problems when performing adaptive interim analyses. It is indeed a fundamental problem to keep integrity and persuasiveness of results from adaptive designs. Contrary to conventional sequential designs, in adaptive interim analysis all relevant accumulating information from inside and outside the trial should be available to support the decisions of a properly selected group of people. This leads to a serious complication of the processes needed to organize such an interim analysis. Specific statistical tools to support decisions are also required. Adaptive designs have an additional fundamental feature of flexibility: They can be started with a conventionally planned group sequential design. At any (unscheduled) time the remainder of this pre-planned design can be replaced by an adapted design which preserves the conditional type I error rate (Müller and Schäfer, 2004). This is equivalent to a continuous application of recursive combination tests where the combination functions are implicitly defined by the pre-planned design. Such designs can be looked at as perfect tool to deal with the unexpected. The price to be paid for such a wide field of flexibility is mainly known. From their perspective the authors do not address possible merits of flexibility but rather target the many problems a regulator has to consider when reviewing an adaptive design application.

A similar position is behind the paper by Koch (2006) who is working as a statistician at the BFARM in Berlin. He starts with examples of why more flexibility is needed in late phase II or phase III trials, covering the options considered in the previous paper and additionally the options of widening the inclusion criteria and changing the dose ore treatment schedule. The motivating example that "increased flexibility need not necessarily be an advantage" deals with the situation that during a trial a non-inferiority margin "may no longer be relevant as the assumptions, under which this margin seemed to be acceptable, are not met in the current trial". This is a general problem which can not be ascribed to flexible designs. Furthermore, I can not see why this is an argument against adaptive strategies which may be helpful for the proper choice of the margin in such a situation. What he calls "The basic identification problem: Interim analysis may endanger the integrity of the trial" is valid in principle for any type of sequential trial. He points out the possible impact of unblinding and summarizes that "a subtle balance should be achieved between the need to access accumulating information and the risk that the integrity of the trial may be compromised after an interim analysis". I can agree with this statement but interim analyses are nowadays not questioned in principle to be a potentially useful tool in clinical trials. The critique of the method of simply combining p-values is again unspecific for flexible designs because test statistics of group sequential designs are essentially also a combination of stage-wise p-values. The author asks for adaptation options to be addressed carefully in the planning phase, which is fully acceptable for those adaptations which can be anticipated (e.g., not all safety issues which might arise can be foreseen).

It is interesting that the author points out that the combination principle "provides additional breakpoints for checks of consistency. If consistency is found, this can increase the credibility of the overall finding". The control of the treatment x stage interaction was an intensively discussed issue even in the early papers on adaptive designs (e.g., Bauer and Köhne, 1994). For sample size calculation the author advocates "adequate discussion of sample size issues at the planning stage" and, if sample size reassessment is applied at all, the use of blinded methods. With regard to modification of the primary endpoint, he prefers to define a multiple test problem with the rival candidates in the planning phase (like the authors of the previous paper). For the widely accepted adaptations of dropping a treatment (e.g., the Placebo group) the author raises the concern that the populations may change if blinding is not maintained for investigators and patients. Predictably, he has reservations about the use of an adaptive design combining phase II and III instead of two independent pivotal trials. For non-inferiority trials he gives the agreeable advice that "it is wise to plan the study as a non-inferiority trial and to foresee in the plan how a switch to superiority could be accomplished based on the results". This is intrinsic to most of such hierarchical test procedures, because non-inferiority is always established when superiority is established but not vice versa.

The author agrees that changing the sample size allocation ratio to the different treatments may have advantages, e.g., to get increased data on an experimental treatment. Note that this is just an example for a specific sample size reassessment rule, just as is dropping a treatment so that its allocation ratio is set to zero.

At the end the author gives a friendly outlook: "The idea of design adaptation has in some instances substantially improved our understanding of clinical trials: ... clinical trials have also been modified in the past. Criteria for inclusion or exclusion have sometimes been reconsidered rather carelessly. ... We strongly believe that adaptive designs have a place in phase III but hope that this place will be explored carefully in order to avoid exaggerated expectations that cannot be fulfilled ...". There is certainly a need for careful exploration when looking at the limited number of published applications in medicine. The average quality of applying and presenting adaptive designs is rather poor (Bauer and Einfalt, 2006).

The three papers cover the wide range of how flexible designs are treated nowadays in clinical trials. It is agreed that new contributions from biostatistics are required, e.g., in the area of registration of innovative therapies (O'Neill, 2006). Such research is currently done, e.g., in the area of adaptive design. On the other hand, it is made clear that these methods have to be applied in a convincing and transparent way to be acceptable in areas with impact on public health (Hung et al., 2006; Koch, 2006). My own experience is that this should be possible. However, generally it will need more planning, input and logistics than conventional designs for clinical trials (including increased involvement of biostatisticians). This has also to be taken into account when weighting other (more methodological) arguments for and against the use of an adaptive design in a specific medical environment.

References

Bauer, P. and Einfalt, J. (2006). Application of adaptive designs – a review. *Biometrical Journal* **48**, 493–506.

Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine*, **18**, 1833–1848.

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

Brannath, W., König, F., and Bauer, P. (2006). Estimation in flexible two-stage designs. *Statistics in Medicine* **25**, in press.

O'Neill, R. T. (2006). FDA's critical path initiative: A perspective on contributions of biostatistics. *Biometrical Journal* **48**, 559–564.

Hung, H. M. J., O'Neill, R. T., Wang, S.-J., and Lawrence, J. (2006). A regulatory view on adaptive/flexible clinical trial design. *Biometrical Journal* **48**, 565–573.

Koch, A. (2006). Confirmatory clinical trials with an adaptive design. *Biometrical Journal* **48**, 574–585.

König, F., Bauer, P., and Brannath W. (2006). An adaptive hierarchical test procedure for selecting safe and efficient treatments. *Biometrical Journal* **48**, 663–678.

Zehetmayer, S., Bauer, P., and Posch, M. (2005). Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics* **21**, 3771–3777.

