Comments from the Joint Working Group "Adaptive Designs and Multiple Testing Procedures" of the Austro-Swiss and the German Region of the IBS

on the Draft Guidance

Adaptive Designs for Clinical Trials of Drugs and Biologics; Draft Guidance for Industry

Docket ID: FDA–2018–D–3124

of the U.S. Food and Drug Administration

Contributions by (in alphabetic order): Werner Brannath, Ekkehard Glimm, Silke Jörgens, Peter Kimani, Johannes Krisam, Tobias Mütze, Deepak Parashar, Nick Parsons, Martin Posch, Rene Schmidt, Nigel Stallard, Susanne Urach.

Line number(s) of the relevant text	**Comment**
Line 58ff.	Definition of the terms “non-comparative” and “comparative” analysis should be sharpened and the advantage over the terms “unblinded” and “blinded” should be discussed. We propose to add the statement from lines 376-378 to the corresponding paragraph in the Important Concepts Section, lines 58-64.
Lines 83-88	It is not sufficiently clear how the concepts of “fixed sample trials” and a “non-adaptive trials” differ. Do the terms coincide or does the definition of fixed sample trials also cover adaptive trials where no adaptions regarding the recruitment and allocation of patients are performed? In addition, the schedule of analyses in time-to-event trials is typically determined in the calendar time or information time scale rather than by sample size. We therefore suggest use of a more concise terminology.
Lines 90-91	The definition of bias is somewhat vague and could be clarified by given examples for statistical parameters that quantify bias, e.g.: “Statistically, bias is quantified via a location parameter of the treatment effect estimate like its expectation or median.”
Lines 254, 323-325, 334, 450-	We believe that there is slightly too much emphasis on “completely pre-specified adaptations”. While for some testing procedures a full pre-

452, 543f, 963-966 and 1072ff.	specification of the adaptation algorithm is required to demonstrate type I error rate control, others, as the p-value combination or the conditional error probability approach allow for more flexibility. For example, the methods of adaptive designs pioneered by Bauer and Köhne (1994) and Proschan and Hunsberger (1995) are intended to allow flexibility without compromising type I error control. Still, pre-planning and guidance to DMCs for decision-making are important, however, even if it is decided in the interim analysis to deviate from pre-specified rules, type I error rate control is still guaranteed. In contrast, methods based on fixed rules (e.g. Stallard and Todd, 2003), typically make explicit use of these rules (e.g. by calculating critical values from integrating over conditional distributions resulting for later stages only under these very specific rules). The latter methods are not often used in practice because of their inflexibility. We believe that the guidance does not intend to promote these strict approaches. We propose to allow for more flexibility for approaches that control the type I error rate also without full pre-specification of the adaptation algorithm. Bauer P & Köhne K (1994). Evaluation of experiments with adaptive interim analyses. Biometrics 50: 1029-1041. Proschan MA & Hunsberger SA (1995). Designed extension of studies based on conditional power. Biometrics 51: 1315-1324. Stallard N & Todd S (2003). Sequential designs for phase III clinical trials incorporating treatment selection. Statistics in Medicine 22(5): 689-703.
Lines 310-317	Bias adjusted estimators often have a higher variance and are less precise. Therefore, measures that take into account both, bias and variance, are important to assess the properties of estimates. For example, biased estimators with a low MSE may be a more useful basis for decision making than unbiased estimators with larger variances.
Lines 378-381	It is important to note that not only in open-label trials, knowledge on treatment assignment may be inferred from the blinded data, e.g., if information on surrogate or safety endpoints are available that are affected by treatment. If this information is used for adaptations without appropriate adjustment, this may lead to biased tests (see, e.g., Posch, M., & Proschan, M. A. (2012). Unplanned adaptations before breaking the blind. Statistics in Medicine, 31(30), 4146-4153. Żebrowska, M., Posch, M., & Magirr, D. (2016). Maximum type I error rate inflation from sample size reassessment when investigators are blind to treatment labels. Statistics in Medicine, 35(12), 1972-1984.)
Lines 381-382	The draft guidance states that “In general, adequately pre-specified adaptations based on non-comparative data have a negligible effect on the Type I error probability.” This, however, cannot be generally stated. For the situation of equivalence and non-inferiority trials, there are practice-relevant situations where the actual type I error rate increased to up to 7% when applying a sample size reassessment based on non-comparative data with a nominal significance level of 5% (c.f. Friede, T., & Kieser, M. (2003). Blinded sample size reassessment in non-inferiority and equivalence trials. Statistics in Medicine, 22(6), 995-1007). This inflation should not be regarded as negligible, and thus one cannot claim that sample size reassessments based on non-comparative data generally

	maintain the type I error level. Before implementing such methods in a trial, the impact on the type I error rate should be thoroughly investigated.
Lines 414-416	This sentence is misleading. It could be interpreted as saying that adaptive designs do not control the type I error rate.
Line 426ff.	If subject accrual is fast compared to the follow-up time of individual patients, the interim analyses are based on only part of the outcome data of the enrolled patients and the complete data will be available only at a later time point. In case of early stopping, these data has to be followed up and included in a final analysis. Guidance on how to deal with such delayed responses and potential reversals of interim decisions would be important (e.g. Faldum, A. & Hommel, G. (2007). Strategies for including patients recruited during interim analysis of clinical trials. Journal of Biopharmaceutical Statistics 17(6), 1211-1225. Hampson, L.V. & Jennison, C. (2013). Group sequential tests for delayed responses. Journal of the Royal Statistical Society: Series B 75(1), 3-39).
Lines 431-434	Although the expected sample size under the alternative for which the trial is powered, is usually lower in group sequential designs, the maximum sample size of the trial is higher. A corresponding note of caution could be added.
Lines 479-484	We believe that the guidance should discuss in more detail the acceptability of “binding” vs “non-binding” futility rules.
Lines 486-494	Although as stated here, in many settings it is not desirable to stop early for efficacy because of the resulting lack of safety data etc., it could be acknowledged that early stopping for efficacy can be appropriate if sufficient information on safety is already available at the time of the interim analysis.
Line 565	“modify cations” should probably be replaced by “modifications”
Lines 606-639	It is not clear in this section when the treatment arms are different doses of the same drug and when they are different drugs. If the issues are identical in these two cases, this should be made clear by adding a sentence to say that different arms could be either different treatments or different doses of the same drug. Alternatively, there could be two subsections, one on Adaptations to Dose Selection (including lines 606-616) and one on Treatment Arm Selection (lines 616-639, while replacing “dose” by “treatment” everywhere).
Lines 691-697	If there is uncertainty about the treatment effects of multiple primary endpoints, a multiple testing procedure can be applied as alternative to adaptive designs. As the multiple testing procedure is based on a larger sample size for each endpoint compared to the adaptive design (which disregards data once an endpoint is dropped in an interim analysis) the multiple testing procedure is typically more efficient. On the other hand, adaptation of endpoints may be appropriate if the measurement of endpoints is burdensome for patients (e.g., because an invasive diagnostic procedure is required) and measurement of the endpoint is discontinued after an endpoint has been dropped.
Lines 726-728	Even if the asymptotic distributions can be derived, these may substantially deviate from the distributions in finite samples and therefore even in these settings simulations may be required to assess the operating

	characteristics. Guidance would be useful, to which extent also finite sample properties should be investigated via simulation. For example, in group sequential designs quantile substitution approaches are applied to derive critical values for t-tests which control the type I error rate only asymptotically. Here simulations can be used to demonstrate that the inflation is negligible unless sample sizes are very low.
Lines 731-733	If adaptive tests based, e.g., on combination tests are applied, even for complex adaptive designs covering several types of adaptation, no simulations are required to demonstrate familywise type I error rate control (see, e.g., Wassmer, G. & Brannath, W. (2016). Group Sequential and Confirmatory Adaptive Designs in 1357 Clinical Trials. Springer series in pharmaceutical statistics, New York: Springer). An example on how these methods guarantee familywise type I error rate control in a rather complex design can be found e.g. in Gutjahr G, Posch M, Brannath W. (2011). Familywise error control in multi-armed response-adaptive two-stage designs. Journal of Biopharmceutical Statistics, 21(4):818-830.
Lines 748-751	It would be important to emphasize here that the simulation scenarios should not only include varying assumptions on nuisance parameters for the primary endpoints, but also on the distribution of all outcomes that are used in the adaptation decisions.
Lines 779- 786	Note that the worst case is not independence, but, e.g., for two multivariate normal endpoints a correlation of -1. Moreover, for the Bonferroni and Holm procedures it is theoretically known that it controls the multiple type I error rate for all correlation structures. This also applies to the weighted Bonferroni tests and all related closed test procedures.
Lines 946-953	The potential bias in treatment effect estimates may result in an underpowered phase III when used for the sample size calculation. This may be mentioned here as well. For instance, the naïve maximum likelihood estimate of the response probability in a Simon design is well known to be upwards biased; see e.g. Chang MN, Wieand HS and Chang VT (1989). The bias of the sample proportion following a group sequential phase II clinical trial. Stat Med; 8: 563–570.