Adaptive Designs and Multiple Testing Procedures Workshop 2021: June 10th – 11th

June 10th, 9:00-13:30 CEST

9:00 - 9:15	**Welcome and Opening** Thomas Asendorf, Tim Friede
Invited Session	Chairs: Frank Bretz, Sonja Zehetmayer
9:15 - 9:45	**How (not) to rescue a trial that has been impacted by the COVID-19 pandemic** Cornelia Ursula Kunz | *Boehringer Ingelheim Pharma GmbH & Co. KG, Germany*
9:45 - 10:15	**Adaptive designs for clinical trials in COVID-19: the RECOVERY-Respiratory Support trial** Nigel Stallard | *Warwick Medical School, United Kingdom*
10:15 - 10:30	Discussion & Coffee Break (Break-Out Rooms)
Session 1	Chairs: Martin Posch, Geraldine Rauch
10:30 - 11:00	**An adaptive design for early clinical development including interim decision for single-arm trial with external controls or randomized trial** Heiko Götte | *Merck Healthcare KGaA, Germany*
11:00 - 11:30	**Binding forces – a conditional performance score and an optimization approach towards adaptive clinical trial designs put together** Carolin Herrmann | *Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Germany*
11:30 - 12:00	**Multivariate Adaptive Group Sequential Survival Tests** Rene Schmidt | *IBKF, WWU Münster, Germany*
12:00 - 12:15	Discussion & Coffee Break (Break-Out Rooms)
Session 2	Chairs: Andreas Faldum, Tobias Mütze
12:15 - 12:45	**On the use of comparison regions in visualizing stochastic uncertainty in two-parameter estimation problems** Werner Vach | *Basel Academy for Quality and Research in Medicine, Switzerland*
12:45 - 13:15	**Uni- and multivariate comparisons of treatment or dose groups versus control or placebo group: closed testing vs. simultaneous procedures** Ludwig A. Hothorn | *Leibniz Uni Hannover, Germany*
13:15 - 13:30	Discussion & Closing

June 11th, 9:15-12:45 CEST

Session 3	Chairs: Lisa Hampson, Carolin Herrmann
9:15 - 9:45	**Blinded Sample Size Re-assessment in multi-arm trials with an upper limit on the maximum sample size** Cornelia Ursula Kunz | *Boehringer Ingelheim Pharma GmbH & Co. KG, Germany*
9:45 - 10:15	**Group sequential methods for the Mann-Whitney parameter** Claus Peter Nowak | *Charité - Universitätsmedizin Berlin, Germany*
10:15 - 10:45	**Blinded sample size recalculation in adaptive enrichment designs** Marius Placzek | *Unversitätsmedizin Göttingen, Germany*
10:45 - 11:00	Discussion & Coffee Break (Break-Out Rooms)
11:00 - 11:30	**Working Group Meeting**
Session 4	Chairs: Anna-Maria Kloidt, Rene Schmidt
11:30 - 12:00	**Online control of the False Discovery Rate in platform trials** Sonja Zehetmayer | *Medical University of Vienna, Austria*
12:00 - 12:30	**Testing procedures for the comparison of multiple characteristics of different survival functions** Robin Ristl | *Medizinische Universität Wien, Austria*
12:30 - 12:45	Discussion & Closing

Registration

Registration to the workshop is possible using the conftool under: https://www.conftool.org/admtp-workshop-2020/index.php

The ADMTP Working Group

The Adaptive Designs and Multiple Testing Procedures working group has the aim to promote the development and application of biostatistical procedures in the fields of multiple testing and group adaptive designs. The working group organises an annual workshop, which offers presentations on state of the art research on adaptive designs and multiple testing procedures and related fields, but also showcases of the practical implementation of these methods. For further information on the working group see: http://www.biometrische-gesellschaft.de/arbeitsgruppen/adaptive-designs-multiple-testing-procedures.html

June 10th, 9:15-9:45

How (not) to rescue a trial that has been impacted by the COVID-19 pandemic

Cornelia Ursula Kunz
Boehringer Ingelheim Pharma GmbH & Co. KG, Germany

The start of the COVID-19 pandemic was also the start for a hunt for treatments, vaccines, and/or diagnostics for the new disease. Some of these trials use novel, efficient trial designs including platform trials and adaptive group-sequential designs. While considerable efforts are being made to set up trials in COVID-19, most of the ongoing trials continue to be in other disease areas. Often these trials had been planned years before the pandemic and started recruitment before COVID-19 existed.
Nevertheless, the pandemic led to several consequences for ongoing trials in non-COVID-19 conditions. To protect patient safety, across the world, clinical trials have been stopped or temporarily paused to possibly restart later. Endpoints had to be changed, treatment regiments needed to be reconsidered. A natural question that came to mind is whether the trial design can or even should be changed to account for some of the problems that have arisen.
The talk mainly focuses on possible design adaptations, in which situations they can be helpful, and how to ensure type I error rate control (if required). In more detail, approaches to resizing a trial affected by the pandemic are developed including considerations to stop a trial early, the use of group-sequential designs or sample size adjustment. All methods considered are implemented in a freely available R shiny app.

June 10th, 9:45-10:15

Adaptive designs for clinical trials in COVID-19: the RECOVERY-Respiratory Support trial

Nigel Stallard
Warwick Medical School, United Kingdom

The SARS-CoV-2 pandemic has led to unprecedented clinical research activity. An important part of this research has focused on randomized controlled clinical trials to evaluate potential therapies for COVID-19. The results from this research need to be obtained as rapidly as possible. This presents a number of challenges associated with considerable uncertainty over the natural history of the disease, the number and characteristics of patients affected, and the emergence of new potential therapies. These challenges make adaptive designs for clinical trials a particularly attractive option. This talk will discuss some of the advantages and challenges of adaptive designs in the pandemic setting illustrated by details of an ongoing trial in COVID-19 respiratory support.

June 10th, 10:30-11:00

An adaptive design for early clinical development including interim decision for single-arm trial with external controls or randomized trial

Heiko Götte[1], Marietta Kirchner[2], Johannes Krisam[2], Arthur Allignol[1], Francois-Xavier Lamy[1], Armin Schüler[1], Meinhard Kieser[2]
[1]Merck Healthcare KGaA, Germany; [2]Institute of Medical Biometry and Informatics, University of Heidelberg, Germany

Keywords: confounder adjustment, preference score, loss function, real-world data, proof-of-concept

In early clinical development, randomized controlled trials (RCT) or single-arm trials with external controls (SATwEC) are design options which allow adjustment for confounding: RCT via design, SATwEC via analysis using propensity score methods. SATwEC requires less investment than RCT. However, if the confounder space substantially differs between experimental and external control group, the SATwEC might lead to inappropriate decisions for further development.
We develop an adaptive two-stage design (ATD) for early clinical development that reduces the risk of unreliable decision making at the end of a SATwEC. In stage I, subjects are solely assigned to the experimental group. If at interim the propensity score distributions of internal and external data are comparable based on the preference score, the subjects in stage II will again be solely assigned to the experimental arm; if not, a randomized stage II will be conducted.
In a simulation study guided by a motivating example, data is generated using a time-to-event model with observable and unobservable confounders. The confounder space is varied to investigate the impact on false go/stop probabilities as well as a loss function which reflects quality of treatment effect estimates and decision making.
The proposed ATD provides a compromise between optimizing quality (as expressed by false go/stop probabilities and the loss function) and investment (defined by sample size and trial duration).

June 10th, 11:00-11:30

Binding forces – a conditional performance score and an optimization approach towards adaptive clinical trial designs put together

Carolin Herrmann[1], Maximilian Pilz[2], Meinhard Kieser[2], Geraldine Rauch[1]
[1]*Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Germany;* [2]*Institute of Medical Biometry and Informatics, University Medical Center Ruprecht-Karls University Heidelberg, Germany*

Keywords: sample size recalculation, performance evaluation, optimization

Adaptive clinical trial designs provide a tool to react to planning uncertainties in clinical trials. They offer the opportunity either to stop a trial early or to adapt the sample size during an ongoing trial. There exist different possibilities to adapt the sample size. Many recalculation rules rely on conditional power arguments [e.g. 1, 2]. Pilz et al. [3] follow an optimization approach, where multiple design parameters (per-stage sample sizes and stopping criteria) are optimized simultaneously. However, an open question is the scoring criterion in the optimization procedure. Herrmann et al. [4] suggested a conditional performance score that focuses on the evaluation of the conditional power and sample size in adaptive study designs with sample size recalculation conditional on not stopping the trial early.

Therefore, it seems natural to use the conditional performance score [4] as scoring criterion in the optimization approach [3]. The underlying idea of the optimization approach was implemented in the R package adoptr [5] for a finite-dimensional parameter space.
In a first analysis step when combining the conditional performance score and the optimization approach, all design parameters apart from the stage two sample size are fixed such that the stage two sample size is optimized with respect to the conditional performance score. In a second step, different prior distributions underlying the power calculations are investigated.

The optimization approach with the conditional performance score as scoring criterion leads often to concave curves for the stage two sample sizes. The underlying prior distribution of the power calculations has a clear influence on the stage two sample sizes and therefore also on the corresponding sample size curves in dependence of the observed interim effect. Overall, continuous prior distributions lead to flatter sample size curves. Moreover, we have seen that the corresponding optimal conditional score value for a given setting is usually smaller than the theoretical best conditional score value of 1.

The optimization approach with respect to the conditional performance score follows two aims: On the one hand, “optimal” stage two sample sizes with respect to the score can be derived. On the other hand, an upper boundary for the conditional performance score in a specific application setting is calculated. This boundary can be used as benchmark when comparing different sample size recalculation approaches with respect to the conditional performance score. However, especially the resulting sample size curves should always be considered with caution.

The R package adoptr can help to determine realistic best conditional score values for specific settings. Moreover, it can be used to derive sample size curves that are optimal with respect to the conditional performance score.

1. Proschan, M. A., & Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. Biometrics, 1315-1324.
2. Mehta, C. R., & Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: a practical guide with examples. Statistics in Medicine, 30(28), 3267-3284.
3. Pilz, M., Kunzmann, K., Herrmann, C., Rauch, G., & Kieser, M. (2019). A variational approach to optimal two-stage designs. Statistics in Medicine, 38(21), 4159-4171.
4. Herrmann, C., Pilz, M., Kieser, M., & Rauch, G. (2020). A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation. Statistics in Medicine, 39(15), 2067-2100.
5. Kunzmann, K. & Pilz, M. (2020). Adoptr. R package version 0.4.1. https://CRAN.R-project.org/package=adoptr.

June 10th, 11:30-12:00

Multivariate Adaptive Group Sequential Survival Tests

Rene Schmidt, Andreas Faldum, Moritz Fabian Danzer
IBKF, WWU Münster, Germany

Keywords: adaptive, multivariate, survival test

Classical adaptive designs for survival trials are commonly restricted to one primary time-to-event endpoint and suffer from limitations regarding the use of information from other endpoints in interim design changes. Here, we discuss designing adaptive group sequential tests for hypotheses on the multivariate survival distribution derived from multi-state models, while making provision for data-dependent design modifications based on all involved time-to-event endpoints. We explicitly illustrate application of the methodology to one-sample tests for the joint distribution of progression-free survival (PFS) and overall survival (OS) in the context of an illness-death model, and discuss the extension to the two-sample setting.

June 10th, 12:15-12:45

On the use of comparison regions in visualizing stochastic uncertainty in two-parameter estimation problems

Werner Vach[1], Maren Eckert[2]
[1]Basel Academy for Quality and Research in Medicine, Switzerland; [2]Institute for Medical Biometry and Statistics, Medical Center & Medical Faculty, University of Freiburg, Germany

Keywords: two-dimensional confidence regions, post hoc tests

When considering simultaneous inference for two parameters, it is very common to visualize stochastic uncertainty by plotting two-dimensional confidence regions. This allows us to test post hoc null hypotheses about a single point in a simple manner. However, in some applications the interest is not in rejecting hypotheses on single points, but in demonstrating evidence for the two parameters to be in a convex subset of the parameter space. The specific convex subset to be considered may vary from one post hoc analysis to another. Then it is of interest to have a visualization allowing to perform corresponding analyses.

One approach is the construction of regions (called alpha-level comparison regions) with the following property: If a region of interest is covering completely the comparison region, then the null hypothesis that the true parameter is outside of the region of interest can be rejected at the level alpha.

In the talk it is shown that such comparison regions can be constructed and that 5%-comparison regions are distinctly smaller than 95%-confidence regions, which are by definition also comparison regions.

The application of comparison regions is illustrated using two examples: a) Comparison of the diagnostic accuracy between two tests. b) Benefit-risk analysis of an intervention.

1. Eckert, M, Vach, W. On the use of comparison regions in visualizing stochastic uncertaint in some two-parameter estimation problems. Biometrical Journal. 2020; 62: 598– 609. https://doi.org/10.1002/bimj.201800232

June 10th, 12:45-13:15

Uni- and multivariate comparisons of treatment or dose groups versus control or placebo group: closed testing vs. simultaneous procedures

Ludwig A. Hothorn
Leibniz Universität Hannover, Germany

Keywords: CTP, Dunnett, Williams

Closed testing procedures with or without order /or monotonicity assumption reveal an alternative to the Dunnett (1955) or Williams (1971) procedures. The conditions, size and power are compared for the common univariate case as well as few multiple correlated, possible different-scaled endpoints. The limitation of CTP are dimension and hard to interpret confidence intervals. The key approaches are the asymptotic maxT-test and the multiple marginal model Pipper et al. (2012).

Using the CRAN package multcomp, motivating pre-clinical and clinical examples are demonstrated.

June 11th, 9:15-9:45

Blinded Sample Size Re-assessment in multi-arm trials with an upper limit on the maximum sample size

Cornelia Ursula Kunz
Boehringer Ingelheim Pharma GmbH & Co. KG, Germany

Keywords: blinded sample size reassessment, multi-arm trials, power

The adequacy of the sample size is of key importance in clinical trials. One the one hand a trial needs to be large enough to have sufficient power for detecting a clinically relevant effect. On the other hand, a trial should not be too large for ethical and economic reasons. However, when planning clinical trials, there is often considerable uncertainty regarding the variability of the primary endpoint, and hence the appropriate sample size. Blinded sample size re-estimation allows estimating these nuisance parameters based on blinded data from the ongoing trial. Based on this estimate, the sample size can then be adjusted.

So far, methods for blinded sample size re-estimation have focused on two-arm trials. While it is known that the naïve variance estimator is overestimating the variance and as a result slightly overpowering the trial, the effect on the estimator as well as the power has not been studied in detail for multi-arm trials. Furthermore, while there is often a limit on the minimum sample size, less attention has been paid to a possible upper limit of the sample size. If there is an upper limit, it is often quite large as for example allowing the total sample size to be twice the originally planned sample size. However, in many trials, doubling the sample size is not feasible for logistical and/or financial reasons.

Here we investigate both, the effect of having more than two arms as well as the effect of an upper limit on the sample size. Results will be illustrated using a real trial example.

June 11th, 9:45-10:15

Group sequential methods for the Mann-Whitney parameter

Claus Peter Nowak[1], Tobias Mütze[2], Konietschke Frank[1]
[1]*Charité - Universitätsmedizin Berlin, Germany;* [2]*Novartis Pharma AG, Basel, Switzerland*

Keywords: error spending, nonparametric relative effect, Wilcoxon-Mann-Whitney test, win odds

Late phase clinical trials are occasionally planned with one or more interim analyses to allow for early termination or adaptation of the study. While extensive theory has been developed for ordered categorical data such as the Wilcoxon-Mann-Whitney test, there has been comparatively little discussion in the group sequential literature on how to provide repeated confidence intervals and simple power formulas to ease sample size determination. Dealing more broadly with the nonparametric Behrens-Fisher problem, we focus on the comparison of two parallel treatment arms and show that the Wilcoxon-Mann-Whitney test, the Brunner-Munzel test, as well as a test procedure based on the log win odds, a modification of the win ratio, asymptotically follow the canonical joint distribution. In addition to developing power formulas based on these results, simulations confirm the adequacy of the proposed methods for a range of scenarios. Lastly, we apply our methodology to the FREEDOMS clinical trial (ClinicalTrials.gov Identifier: NCT00289978) in patients with relapse-remitting multiple sclerosis.

June 11th, 10:15-10:45

Blinded sample size recalculation in adaptive enrichment designs

Marius Placzek, Tim Friede
Universitätsmedizin Göttingen, Germany

Incorporating multiple populations and hypotheses in design and analysis plan, adaptive designs promise flexibility and efficiency in trials with subgroup analyses. Adaptations include (unblinded) interim analyses or blinded sample size reviews. An interim analysis offers the possibility to select promising subgroups and reallocate sample size in further stages. Trials with these features are known as adaptive enrichment designs. Such complex designs comprise many nuisance parameters, such as prevalences of the subgroups and variances of the outcomes in the subgroups. Additionally, a number of design options including the timepoint of the sample size review and timepoint of the interim analysis have to be selected. Here, for normally distributed endpoints, a strategy combining blinded sample size recalculation and adaptive enrichment at an interim analysis is proposed, i.e. at an early timepoint nuisance parameters are reestimated and the sample size is adjusted while subgroup selection and enrichment is performed later. Implications of different scenarios concerning the variances as well as the timepoints of blinded review and interim analysis are discussed and design characteristics are presented via simulations.

June 11th, 11:30-12:00

Online control of the False Discovery Rate in platform trials

Sonja Zehetmayer, Martin Posch, Franz König
Medical University of Vienna, Austria

Keywords: platform trials, multiple testing, online False Discovery Rate, group sequential trial

When testing multiple hypotheses, the control of a suitable error rate is desirable even in exploratory trials. For such applications, e.g., the control of the False Discovery Rate (FDR) has been proposed. The FDR is defined as the expected proportion of false positive rejections among all rejected hypotheses. Conventional methods to control the FDR, e.g., the Benjamini-Hochberg procedure, assume that all p-values are available at the time point of test decision. In perpetual platform trials treatment arms can enter and leave the trial at any time during its conduct, i.e, the number of hypotheses is not fixed in advance and the hypotheses are not tested at once, but sequentially. Recently, the concept of online control of the FDR was introduced (Javanmard and Montanari, 2015, 2018, Wason and Robertson, 2020), where hypothesis tests and test decisions can be performed in a sequential manner.

We investigated different procedures to control the online FDR in the setting of platform trials. The results hugely depend on the prior distribution of effects sizes, e.g., whether the true alternatives are uniformly distributed over time or not. We further investigated the impact of design parameters on operating characteristics such as the overall power, which is the proportion of rejected alternatives among all alternatives. Furthermore, we investigated the impact of including both concurrent and non-concurrent control data on error rates and overall power. By including the latter the power can be increased, but the control of the FDR is negatively affected in case of time trends. Finally, we show how the procedures have to be modified to allow for interim analyses with the option of early stopping for individual hypotheses.

June 11th, 12:00-12:30

Testing procedures for the comparison of multiple characteristics of different survival functions

Robin Ristl[1], Heiko Götte[2], Armin Schüler[2], Martin Posch[1], Franz König[1]
[1]*Medizinische Universität Wien, Austria;* [2]*Merck KGaA, Germany*

Keywords: non-proportional hazards, counting process, simultaneous confidence intervals, multiple testing procedure

Survival and other time-to-event variables are widely used endpoints in clinical trials. In a typical clincial trial comparing a treatment and a control group with respect to a time-to-event outcome, confirmatory inference and quantification of the treatment effect is based on the hazard ratio estimated from a proportional hazards model. In absence of the proportional hazards assumption, the usual hazard ratio estimate is not a reliable measure of treatment effect, though, because it depends on the censoring pattern, the study duration and the recruitment regimen in combination with the actual survival distribution. In cases of crossing hazard functions or crossing survival functions the use of the hazard ratio as effect measure is further reduced.

In these settings it may be more appropriate to quantify the difference in survival functions by a set of more than one, well interpretable, parameters. Such parameters may be the difference in predefined x-year survival probabilities, e.g. in 1-year and 2-year survival, the difference in quantiles of the survival functions, e.g. difference in medians, an average hazard ratio or the difference in restricted mean survival time up to a preset time-point.

Whenever more than one parameter is considered to assess the treatment effect in a confirmatory way, an inference approach with control of the family wise type I error rate and predefined simultaneous coverage of confidence intervals is warranted. By applying the counting process representation of survival function estimates, we show that the proposed estimates are asymptotically multivariate normal and we derive an estimate of their asymptotic covariance matrix and normality-based simultaneous confidence intervals. As an alternative method for constructing simultaneous confidence intervals we apply the perturbation approach for survival function estimates, which is similar to a parametric bootstrap.

The finite sample properties of the proposed methods are investigated in a simulation study showing coverage probabilities close to the nominal value even for moderate sample sizes.