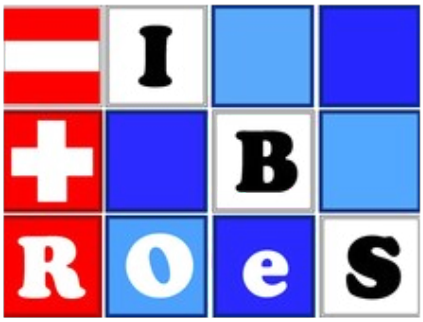
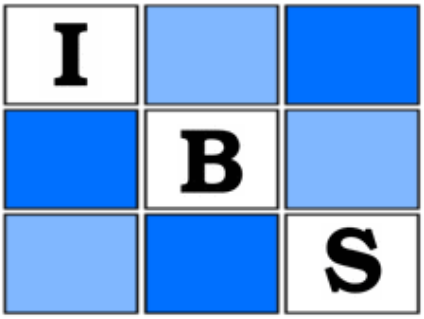
ADAPTIVE DESIGNS AND MULTIPLE TESTING PROCEDURES

Workshop October 4 - 5, 2018

in Bremen





Deutsche Region

Universität Bremen

Kompetenzzentrum für Klinische Studien Bremen

Organization comittee:

Werner Brannnath (Bremen), Thorsten Dickhaus (Bremen),
Florian Klinglmüller (Wien) and Rene Schmidt (Münster)

Sponsored by:

GCP-Service International Ltd. & Co. KG

medac GmbH

Novartis Pharma AG

F. Hoffmann-La Roche Ltd.

LOCATION

The conference will take place inside the building “Geisteswissenschaften 2“, shortly named „GW2“ in the rooms B3009 and B3010.

Universität Bremen
GW2
Enrique-Schmidt Straße
28359 Bremen - Germany

Arrival by car Arrival from north/south of Germany via expressway A1 or A27: Comming from the expressway A1, exit at the junction „Bremer Kreuz“ to the expressway A27 in the direction to „Bremerhaven“. On the A27, take the exit 19 „Horn-Lehe/Universität“ in the direction „Universität“. Turn right at the second traffic light into „Universitätsallee“. After 150 m turn right into „Enrique-Schmidt-Straße“. There you'll find various parking lots. The GW2 is located on the left side of „Enrique-Schmidt-Straße“.

From Airport or Main Station: Outside the airport or in front of the main station (tram platform E) take the tram line 6 in the direction of „Universität-Nord“ to „Universität/Zentralbereich“. You can buy a ticket in the tram at a ticket machine (2,80 € per person). When you exit the tram, walk straight ahead through the glass hall (we will have signs guiding you). Behind the glass hall you will find the GW2.

Map of University Bremen

CONFERENCE DINNER

There will be a conference dinner on

Thursday, October 4, 19:30

at the historical „Ratskeller“ in the Center of Bremen. Registration for the conference dinner is required.

The address is:

Bremer Ratskeller
Am Markt
28195 Bremen

SCIENTIFIC PROGRAM OVERVIEW

Thursday morning, October 4

08:00	Registration
08:40 - 08:50	Welcome
08:50 - 10:30	Session 1: Designs with unblinded sample size recalculation
10:30 - 11:00	Coffee break and Poster Session
11:00 - 12:40	Session 2: Advanced multiple testing and adaptive design methodology I
12:40-14:00	Lunch time

Thursday afternoon, October 4

14:00 - 15:40	Session 3: Subgroup and biomarker analysis
15:40 - 16:10	Coffee break and Poster Session
16:10 - 17:50	Session 4: Memorial session in honor of Willi Maurer
18:00	Meeting of the IBS-DR/ROeS Working Group on „Adaptive Designs and Multiple Testing Procedures“
19:30	Conference dinner at the Ratskeller Bremen

Friday morning, October 5

08:20 - 10:00	Session 5: Adaptive designs and multiple testing in pharmaceutical research
10:00 - 10:30	Coffee break and Poster Session
10:30 - 12:10	Session 6: High dimensional multiple testing
12:10 - 13:30	Lunch time

Friday afternoon, October 5

13:30 - 15:10	Session 7: Advanced adaptive design and multiple testing methodology II
15:10 - 15:40	Coffee break and Poster Session
15:40 - 17:20	Session 8: Blinded sample size reviews and software for adaptive designs
17:20	End of workshop

SCIENTIFIC PROGRAM - DETAILED TIME SCHEDULE

Session 1: Designs with unblinded sample size recalculation
Chairs: Silke Jörgens, Guido Knapp

Thursday, 08:50 - 10:30

08:50 - 09:15

Johannes Krisam (University of Heidelberg), Dorothea Weber, Richard F. Schlenk, Meinhard Kieser: *Including matched control patients in single-arm adaptive two-stage phase II trials – the Matched-Threshold-Crossing (MTC) design* (p. 21)

09:15 - 09:40

Bart Michiels (Johnson& Johnson), Wilbert van Duijnhoven: *An adaptive design for a self-limiting disease* (p. 27)

09:40 - 10:05

Carolin Herrmann (Charité - Universitätsmedizin Berlin, Berlin Institute of Health), Meinhard Kieser, Maximilian Pilz, Kevin Kunzmann, Geraldine Rauch: *A new conditional performance score for evaluating sample size recalculation rules in adaptive designs* (p. 18)

10:05 - 10:30

Geraldine Rauch (Charité - Universitätsmedizin Berlin, Berlin Institute of Health), Meinhard Kieser, Maximilian Pilz, Kevin Kunzmann, Carolin Herrmann: *A new semi-Bayesian rule for sample size recalculation in adaptive designs* (p. 34)

Session 2: Advanced multiple testing and adaptive design methodology I
Chairs: Thorsten Dickhaus, Rene Schmidt

Thursday, 11:00 - 12:40

11:00 - 11:25

Helmut Finner (German Diabetes Center), Markus Roters: *Probability inequalities between one- and two-sided union-intersection tests* (p. 13)

11:25 - 11:50

Susanne Urach (Medical University of Vienna), Franz König, Martin Posch: *Testing endpoints with unknown correlation* (p. 39)

11:50 - 12:15

Marcus Vollmer (University Medicine Greifswald): *Estimation of Sample Size and Power for Dunnett's Testing Setups with Unequal Effect Sizes* (p. 41)

12:15 - 12:40

Michael Grayling (University of Cambridge), James Wason, Adrian Mander: *Efficient determination of optimized multi-arm multi-stage experimental designs with control of generalized error rates* (p. 17)

Session 3: Subgroup and biomarkers analysis
Chairs: Frank Bretz, Cornelia Ursula Kunz

Thursday, 14:00 - 15:40

14:00 - 14:25

Kaspar Rufibach (Roche), Marcel Wolbers, Ke Li: *More efficient treatment effect estimation in pre-specified subgroups displayed in forest plots for time-to-event outcomes* (p. 36)

14:25 - 14:50

Marcel Wolbers (Roche), Kaspar Rufibach: *Pre-planned subgroup analysis within a group sequential design for a time-to-event endpoint* (p. 43)

14:50 - 15:15

Leandro Garcia Barrado (Hasselt University), Tomasz Burzykowski: *The effect of design-related decisions on operational characteristics of trials that use Bayesian biomarker-driven outcome-adaptive randomization* (p. 14)

15:15 - 15:40

Alexandra Graf (Medical University of Vienna): Testing procedures for confirmatory subgroup analysis based on a continuous biomarker (p. 16)

Session 4: Memorial session in honor of Willi Maurer
Chair: Werner Brannath

Thursday, 16:10 - 17:50

16:10 - 16:35

Willi Maurer, Frank Bretz (Novartis), Xiaolei Xun: *Optimal test procedures for multiple hypotheses controlling the familywise expected loss* (p. 25)

16:35 - 17:00

Ekkehard Glimm (Novartis), Angelka Caputo, Mauritz Bezuidenhoudt: *Testing strategies for group-sequential clinical trials with adaptive dose selection and multiple endpoints* (p. 15)

17:00 - 17:25

Martin Posch (Medical University of Vienna): *Mastering Multiplicity in Clinical Trials: Shortcuts, Graphs and Error Rates* (p. 33)

17:25 - 17:50

Gerhard Hommel (University of Mainz): *Reminiscences of my visits at the ROeS seminars* (p. 19)

Session 5: Adaptive and multiple testing in pharmaceutical research
Chairs: Ekkehard Glimm, Kaspar Rufibach

Friday, 08:20 - 10:00

08:20 - 08:45

Arsénio Nhacolo (University Bremen), Werner Brannath: *Oncology Phase II Adaptive Designs - Treatment effect estimates and their use in planning Phase III trials* (p. 31)

08:45 - 09:10

Saswati Saha (University Bremen), Werner Brannath, Bjoern Bornkamp: *Multiple testing approaches in dose combination trial* (p. 37)

09:10 - 09:35

Benjamin Lang (Boehringer Ingelheim), Cornelia Ursula Kunz: *Adaptive Dose-selection in Equivalence Trials* (p. 24)

09:35 - 10:00

Tobias Mielke (Janessen), Franz König: *Adaptive MCPMod Testing* (p. 28)

Session 6: High dimensional multiple testing
Chairs: Gerhard Hommel, Helmut Finner

Friday, 10:30 - 12:10

10:30 - 10:55

Arnorld Janssen (Heinrich-Heine-University Düsseldorf), Marc Ditzhaus: *Valid and consistent adaptive multiple tests* (p. 20)

10:55 - 11:20

Djalel-Eddine Meskaldji (EPFL), Stephan Morgenthaler: *Moderating the trade-off between type I and type II errors via the scaled false discovery rate* (p. 26)

11:20 - 11:45

David Robertson (University of Cambridge), James Wason: *Online control of the false discovery rate in biomedical research* (p. 35)

11:45 - 12:10

Andre Neumann (University Bremen), Taras Bodnar, Thorsten Dickhaus: *Estimating the proportion of true null hypotheses under arbitrary dependency* (p. 30)

Session 7: Advanced adaptive design and multiple testing methodology II
Chairs: Andreas Faldum, Martin Posch

Friday, 13:30 - 15:10

13:30 - 13:55

Rene Schmidt (WWU Münster), Prof. Dr. Andreas Faldum: *Analysis strategies for adaptive survival trials with multiple time-to-event endpoints* (p. 38)

13:55 - 14:20

Kelly Van Lancker (Ghent University), An Vandebosch and Stijn Vansteelandt: Improving interim decisions in randomized trials by exploiting information on short-term outcomes and prognostic baseline covariates (p. 40)

14:20 - 14:45

Diaa Al Mohamad (Leiden University Medical Center), Jelle Goeman, Erik van Zwet, Eric Cator and Aldo Solari: *Adaptive constrained likelihood ratio testing with application to simultaneous confidence intervals for ranks* (p. 10)

14:45 - 15:10

Sonja Zehetmayer (Medical University of Vienna): *A new omnibus test for the global null hypothesis* (p. 44)

Session 8: Blinded sample size reviews and software for adaptive designs
Chairs: Alexandra Graf, Florian Klinglmüller

Friday, 15:40 - 17:20

15:40 - 16:05

Cornelia Ursula Kunz (Boehringer Ingelheim): *The effect of an upper limit for the sample size in designs with blinded sample size re-assessment* (p. 23)

16:05 - 16:30

Tobias Mütze (Novartis), Salem, Susanna; Benda, Norbert; Schmidli, Heinz; Friede, Tim: *Blinded continuous information monitoring of recurrent events endpoints with time trends* (p. 29)

16:30 - 16:55

Thomas Asendorf (University of Göttingen), Robin Henderson, Heinz Schmidli, Tim Friede: *Blinded Sample Size Reestimation for Time Dependent Negative Binomial Counts: An example in MS and considerations of small samples* (p. 12)

16:55 - 17:20

Gernot Wassmer (University of Cologne), Friedrich Pahlke: *RPACT: An R Program for Confirmatory Adaptive Group Sequential Designs* (p. 42)

ABSTRACTS (TALKS)

Adaptive constrained likelihood ratio testing with application to simultaneous confidence intervals for ranks

Diaa Al Mohamad[1], Jelle Goeman, Erik van Zwet, Eric Cator and Aldo Solari

[1])Leiden University Medical Center,Biomedical data sciences
d.al_mohamad@lumc.nl

We present a new way for constrained likelihood ratio testing. In the literature, the constrained likelihood ratio is tested against the quantile of a mixture of chi-squares with weights that are very difficult to calculate. We propose to use a quantile of only one chi-square with data-dependent degrees of freedom. We prove that the new test has a valid $\alpha-$level. The new test is easy to implement and does not require the calculation of any weights. Moreover, it has more power for alternatives that are close to the null in the sense that few constraints are violated. We use the new test to calculate simultaneous confidence intervals (SCI) for ranks. Rankings of institutions such as schools or hospitals are published regularly in newspapers and journals. These ranks are only estimates of the true ranks based on some sample, and thus bear uncertainty. SCI for ranks can be obtained by dividing the space of parameters into disjoint partitions which are defined through sets of constraints. Then, we apply a likelihood ratio test on each of these partitions and use our new approach. Although the resulting testing problem is very complex, we show that only partitions where the observations agree with the constraints defining these partitions are needed. More interestingly, our new approach is more powerful than the classical approach especially on these partitions which results in a gain both in power and execution time. We present results on a dataset of Dutch hospitals and show the simultaneous CIs for their ranks.

Blinded Sample Size Reestimation for Time Dependent Negative Binomial Counts: An example in MS and considerations of small samples

Thomas Asendorf[1], Robin Henderson, Heinz Schmidli, Tim Friede

[1])Universitätsmedizin Göttingen, Institut für medizinische Statistik
thomas.asendorf@med.uni-goettingen.de

Sample size determination in planning clinical trials strongly depends on prior knowledge of nuisance parameters. Assumptions made on nuisance parameters may be inaccurate for a variety of reasons. Blinded sample size reestimation procedures allow for a recalculation of the sample size within an ongoing trial, by estimating nuisance parameters from accumulated data without unblinding (1). We consider modelling time dependent negative binomial count data, as observed e.g. in multiple sclerosis (MS) trials, using a Gamma frailty model (2,3). Procedures for small-sample statistical inference, sample size estimation and blinded sample size reestimation are derived within this model. Possibilities of incorporating time trends are illustrated and demonstrated on two different trends. A simulation study is conducted to assess the finite sample properties of the procedure and the procedure is demonstrated on an example from MS (4), using the R package spass.

Keywords: Longitudinal Count Data, Sample Size Reestimation, Multiple Sclerosis, Adaptive Designs

References

[1] Friede, Schmidli (2010). "Blinded sample size reestimation with count data: Methods and applications in multiple sclerosis". Stat Med, Vol. 29, pp. 1145-1156

[2] Henderson, Shimakura (2003) "A serially correlated gamma frailty model for longitudinal count data". Biometrika, Vol. 90, pp. 355-366

[3] Fiocco et al. (2009). "A new serially correlated gamma-frailty process for longitudinal count data". Biostatistics, Vol. 10, pp. 245-257

[4] Fernandez et al. (2018). "Adipose-derived mesenchymal stem cells (AdMSC) for the treatment of secondary-progressive multiple sclerosis: A triple blinded, placebo controlled, randomized phase I/II safety and feasibility study". PLOS ONE 13(5): e0195891.

Probability inequalities between one- and two-sided union-intersection tests

Helmut Finner[1], Markus Roters

[1])German Diabetes Center (DDZ), Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany; Institute for Biometrics and Epidemiology
finner@ddz.uni-duesseldorf.de

In this talk we focus on collections of real valued random variables $X = (X_t : t \in T)$ and the validity of the probability inequality

$$P^X(A) \leq P^X(A_1)P^X(A_2), \tag{1}$$

where $A = A_1 \cap A_2$ for suitable sets A_1, A_2. For example, one may think of A_1 and A_2 as lower and upper acceptance regions of multivariate one-sided tests, that is

$$A_1 = \{x \in I\!R^n : x_i \leq d_i,\ i = 1, \ldots, n\},$$
$$A_2 = \{x \in I\!R^n : x_i \geq c_i,\ i = 1, \ldots, n\}.$$

Note that we always have a lower (Bonferroni) bound for $P^X(A)$, that is

$$P^X(A) \geq P^X(A_1) + P^X(A_2) - 1.$$

Already 1939, Wald and Wolfowitz conjectured that (1) should be true for one and two-sided acceptance regions of Kolmogorov-Smirnov tests, cf. [1]. Twenty–eight years later, Vandewiele and Noé confirmed this conjecture and proved several inequalities for Kolmogorov-Smirnov type statistics, cf. [2]. At the same time, Esary, Proschan and Walkup delivered a general theory on the association of random variables in their pathbreaking paper [3]. Herewith it is easy to prove that (1) is valid if X is positively associated and if A_1 is a non-decreasing set and A_2 is a non-increasing set. We illustrate this result by means of one- and two-sided multivariate union-intersection tests based on associated random variables. An important consequence of inequality (1) is that the combination of two one-sided multivariate level $\alpha/2$ tests yields a two-sided test which is only slightly conservative for conventional values of α.

References

[1] Wald, A., Wolfowitz, J. (1939) Confidence limits for continuous distribution functions. Ann. Math. Statist. 10: 105-118.

[2] Vandewiele, G., Noé, M. (1967) An inequality concerning tests of fit of the Kolmogorov-Smirnov type. Ann. Math. Statist. 38: 1240-1244.

[3] Esary, J. D., Proschan, F., Walkup, D. W. (1967) Association of random variables, with applications. Ann. Math. Statist. 38: 1466-1474.

The effect of design-related decisions on operational characteristics of trials that use Bayesian biomarker-driven outcome-adaptive randomization.

Leandro Garcia Barrado[1], Tomasz Burzykowski

[1])Hasselt University (Belgium), The Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)
leandro.garciabarrado@uhasselt.be

Bayesian biomarker-driven outcome-adaptive randomization (OAR) designs have drawn a lot of attention in, e.g., cancer clinical trials. They extend traditional fixed randomization ratio designs by allowing the ratio to change continuously, within the strata defined by biomarker values, based on the collected outcome information. It has been advocated that the adaptation allows simultaneous identification of predictive markers and marker-specific treatments. Despite increasing use of these designs, questions regarding their implementation and operational characteristics are still raised. As an example, we consider a design proposed by Barry et al. (2015). It applies an hierarchical probit model to estimate efficacy of two treatments in two biomarker strata using a binary clinical outcome. The design is characterised by a stopping rule with irreversible suspension of accrual to inefficacious treatment-stratum combinations. OAR is initiated after an initial series of n_0 patients that are randomized according to a 1:1 randomization ratio. In the aforementioned design, one has to make several decisions regarding: criteria for testing futility and efficacy of treatments; the timing (n_0) of the start of OAR; prior distributions to be used; the particularities of Bayesian estimation such as the number of burn-in and posterior iterations, convergence monitoring, etc. We are interested in the influence of the different choices on the operational characteristics of the trial. It appears that some choices have important, and sometimes unexpected, consequences. For instance, using different thresholds for the treatment effect in the criteria for testing of futility and efficacy may lead to counterintuitive results in terms of the sample size requirements for the trial. Care is also needed when deciding about the number of MCMC sampling iterations used in the Bayesian estimation algorithm, because an unexpectedly large number of iterations may be required for some of the parameters of the hierarchical model. Obviously, specification of prior distributions requires thought, because some choices lead to an excessive "borrowing" of information about treatment effect across strata, causing error in conclusions regarding efficacy and/or futility. In our paper, we illustrate and discuss these and other consequences of the choices regarding the design of a Bayesian biomarker-driven OAR trial.

Testing strategies for group-sequential clinical trials with adaptive dose selection and multiple endpoints

E. Glimm[1], A. Caputo and W. Maurer

[1]) Novatis Pharma AG
ekkehard.glimm@novartis.com

The talk discusses the design of a complex clinical trial with several sources of multiplicity:

1. multiple doses of the experimental treatment that are compared to a reference treatment,
2. multiple interim analyses with the potential discontinuation of some treatment arms and
3. different endpoints characterizing treatment success.

It is illustrated how these multiplicities can be dealt with by means of the closed test principle, methods from group sequential testing and combination test methodology and how these elements can be combined to yield an approach that achieves high power while controlling the familywise error rate (FWER). We also discuss how knowledge of the correlation between some of the involved test statistics can be exploited and how to select weights for trial stages in the combination test procedure. We illustrate the use of this design with a trial of a BACE-inhibitor used in Alzheimer's disease. Rejection probabilities under important alternatives to the null hypothesis of no drug effect are investigated analytically and by simulation.

References

[1] E.Glimm, M. Bezuidenhoudt, A. Caputo and W. Maurer (2018): A testing strategy with adaptive dose selection and two endpoints. Statistics in Biopharmaceutical Research 10, 196-203.

Testing procedures for confirmatory subgroup analysis based on a continuous biomarker

Alexandra Graf[1]

[1]) Medical University of Vienna, Center for Medical Statistics, Informatics and Intelligent Systems
alexandra.graf@meduniwien.ac.at

With the advent of personalized medicine, clinical trials studying treatment effects in subpopulations are receiving increasing attention. The objectives of such studies are, besides demonstrating a treatment effect in the overall population, to identify subgroups, based on biomarkers, where the treatment has a positive effect. E.g., for patients with depression, there is a large discussion whether biomarkers have an influence on the outcome of treatments in patients with depression. Although a number of treatment options for such patients are available, no single treatment is universally effective. Continuous biomarkers are typically dichotomized based on thresholds to define two subpopulations with low and high biomarker levels. If there is insufficient information on the dependence structure of the outcome on the biomarker, several thresholds may be investigated. The nested structure of the resulting subgroup test statistics is similar to the structure of the sequence of cumulative test statistics in group sequential trials. Due to the impact of potential prognostic effects of the biomarker, group sequential boundaries may not guarantee control of the family-wise type 1 error rate. We consider the problem of how to design a trial with multiple nested subgroups and optimize the number and choice of candidate thresholds as well as the multiple testing procedure.

Efficient determination of optimized multi-arm multi-stage experimental designs with control of generalized error rates

Michael Grayling[1], James Wason, Adrian Mander

[1])University of Cambridge, MRC Biostatistics Unit
mjg211@cam.ac.uk

Primarily motivated by the drug development process, several publications have now presented methodology for the design of multi-arm multi-stage experiments with normally distributed outcome variables of known variance. Here, we discuss an extension to these past considerations to allow for the design of what we refer to as abcd multi-arm multi-stage experiments. Precisely, we outline a proof of how strong control of the a-generalized type-I familywise error rate can be ensured. We then describe how to attain the power to reject at least b out of c false hypotheses, which is related to controlling the b-generalized type-II familywise error rate. Following this, we detail how a design can be optimized for a scenario in which rejection of any d null hypotheses will bring about termination of the experiment. We achieve this by using a novel, highly computationally efficient, approach for evaluating the performance of a candidate design. Finally, using a real clinical trial as a motivating example, we describe the effect of the design's control parameters on the statistical operating characteristics.

A new conditional performance score for evaluating sample size recalculation rules in adaptive designs

Carolin Herrmann[1], Meinhard Kieser, Maximilian Pilz, Kevin Kunzmann, Geraldine Rauch

[1]) Charité – Universitätsmedizin Berlin, and Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology
carolin.herrmann@charite.de

A precise sample size calculation is of major importance for a successful and efficient clinical trial. Under- or overpowering trials should be avoided for ethical and economic reasons. As calculation of the "correct" sample size in the planning stage is based on a number of parameter assumptions, which are related to a certain level of uncertainty, an adjustment of the sample size during an ongoing trial is appealing. After recruiting and evaluating a first sequence of patients, updated knowledge on the required parameters is available which can be used to adapt the sample size or to decide on an early stopping.
So far, there exist no unique standards to assess the performance of adaptive sample size recalculation rules. Consequently, a fair comparison between different recalculation rules is difficult. Single performance criteria commonly reported are given by the power and the average sample size (under the null- or alternative hypothesis) which are obviously highly correlated. Other performance measures such as the variability of the recalculated sample size and the conditional power distribution are often ignored. Liu et al. [1] were the first who presented a performance score for adaptive designs based on sample size and power criteria. This score compares the power and the average sample size of an adaptive design in relation to the "perfect" fixed design (under the true parameter setting) as a gold standard. The performance score has the potential shortcoming that it does not take into account the variability of sample size and that it is not well defined under the null hypothesis of the underlying test problem. Moreover, it is highly questionable whether the "perfect" fixed sample size design is really a valid gold standard.
Therefore, the need for an optimized performance score combining all relevant performance criteria is evident.
In this talk, we present a new conditional performance score and compare it to the one by Liu et al. [1] for a number of well-known sample size recalculation rules.

References

[1] Liu GF, Zhu GR, Cui L. Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. Stat. Med. 2008, 27:584-596.

Reminiscences of my visits at the ROeS seminars

Gerhard Hommel[1]

[1]) University of Mainz
gerhard.hommel@unimedizin-mainz.de

In 1977, I visited the first time a seminar of the ROeS (Austro-Swiss Region of the IBS), and I continued my visits of this biennial event regularly in the next years. In particular, the tutorial character of the seminars was an important aspect for young scientists. During the seminars, I enjoyed several presentations of Willi Maurer, and in 1985 we started a scientific cooperation. In 1987, Willi Maurer gave an acclaimed talk at the ROeS meeting in Locarno. This was the origin for fixed hypotheses sequence testing and gatekeeping procedures. Later on, I had a fruitful cooperation with Frank Bretz and Willi Maurer, motivated by planning of some study protocols at Novartis. The result were the "consonant weighted Bonferroni procedures" (2007) and, as a consequence, graphical approaches to multiple comparison tests (Bretz, Maurer, Brannath, Posch, 2009).

Valid and consistent adaptive multiple tests

Arnorld Janssen[1], Marc Ditzhaus

[1]) Heinrich-Heine-University Düsseldorf, Mathematical Institute
janssena@math.uni-duesseldorf.de

Simultaneous hypotheses testing for "big data" sets is a very difficult affair. The talk introduces first the modern concept of multiple testing and examples are illustrated. Then new results are presented. The pioneer multiple test of Benjamini and Hochberg (1995) with up to date more than 42000 citations is a basic tool in high dimensional data analysis, for instance in genomics when a huge amount of tests are carried out simultaneously for the same data set. Their test and also improved data dependent adaptive tests of Storey, Taylor and Sigmund (2004) control the so called FDR, see also Heesen and Janssen (2016) for more general adaptive procedures. The FDR is the expectation of the ratio of the number of false rejections and all rejections. Although the FDR can be controlled by some given level alpha the "false discovery proportion" (FDP) may have stochastic fluctuations. In this talk we discuss the consistency for general adaptive multiple tests. We present finite sample and asymptotic results in order to bound deviations of the FDP from the present FDR level.

References

[1] Benjamini and Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B 57(1), 289–300.

[2] Heesen, P. and Janssen, A. (2016). Dynamic adaptive multiple tests with finite sample FDR control. J. Statist. Plann. Inference 168, 38–51.

[3] Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 66(1), 187–205.

Including matched control patients in single-arm adaptive two-stage phase II trials – the Matched-Threshold-Crossing (MTC) design

Johannes Krisam[1], Dorothea Weber, Richard F. Schlenk, Meinhard Kieser

[1]) University of Heidelberg, Institute of Medical Biometry and Informatics
krisam@imbi.uni-heidelberg.de

When a phase I trial has been successfully completed, there are several options regarding the design of the subsequent phase II trial. One can e.g. conduct a single-arm trial, where the response rate in the intervention group is compared to a pre-fixed value for the proportion. As a second option, one can conduct a randomised phase II trial comparing the new treatment with placebo or the current standard. Nonetheless, a problem arises in both approaches when the investigated patient population is very heterogeneous regarding prognostic and predictive factors associated with the response, which is frequently the case, e.g. in oncology. Especially for small sample sizes, the observed response rates may substantially differ from the true response rate since the study population might not well reflect the characteristics of the underlying patient population. Additionally, in a usually small-sized randomised phase II trial, an imbalanced distribution of confounders across treatment arms may cause biased treatment effect estimates, as pointed out by Gan et al. [1]. An adjustment can only be performed for known confounders and may be impeded as imbalanced populations may cause instability in statistical models.
For the situation that a substantial dataset of historical controls exists, which is the case in many clinical fields by use of, e.g. registry data, we propose an approach to enhance the classic single-arm trial design by including matched control patients. This approach overcomes the previously described disadvantages, since the expected outcome of the observed study population can be adjusted based on the matched controls with a comparable distribution of known prognostic and predictive factors and balanced treatment groups lead to stable statistical models. The success of a trial within the proposed design can either be defined by a significant hypothesis test comparing treatment and control group at a specified significance level α, or by a successful crossing of a pre-defined threshold as proposed by Eichler et al [2]. A priori unknown parameters in such a design are the matching rate and the number of matched controls per patient in the intervention group, which, however, can be determined at an interim analysis using an iterative procedure. We propose an adaptive two-stage design with a possible (non-binding) stop for futility in case the observed treatment effect does not seem promising at interim. Furthermore, a sample size recalculation is performed using a conditional power argument taking the matching rate and observed treatment effect into account. Our frequentist approach is unique and novel in the sense that, on the one hand, it allows to incorporate matched historical controls within a two-stage single-arm trial ensuring perfectly stable statistical models and, on the other hand, to deal with the uncertainty about

trial parameters by means of an interim sample size reassessment. Performance characteristics of the proposed two-stage-design are investigated in comprehensive simulation studies. Our proposed methods are illustrated by a real clinical trial example from oncology.

References

[1] Gan et al. (2010) Randomized phase II trials: inevitable or inadvisable? J Clin Oncol 28:2641-2647

[2] Eichler et al. (2016) "Threshold-crossing": A useful way to establish the counterfactual in clinical trials? Clin Pharmacol Ther 100:699-712.

The effect of an upper limit for the sample size in designs with blinded sample size re-assessment

Cornelia Ursula Kunz[1]

[1]) Boehringer Ingelheim
cornelia_ursula.kunz@boehringer-ingelheim.com

Sample size determination is a key issue in the planning phase of a trial. 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.
The sample size is influenced by several parameters including the assumed treatment effect, the type I error, the power, and the variability. While the assumed treatment effect can be specified by medical experts and there is a consensus about the values to be used for type I error and power, there is often uncertainty about the variability affecting the target variable and hence, the sample size cannot be exactly determined either.
A commonly used approach is to base the initial sample size determination on a variability estimate and complement this with a pre-specified blinded sample size re-estimation (bSSR). Based on the updated variability estimate from the bSSR, the sample size is updated to ensure the planned power for the trial analysis. So far, literature on bSSR has focused on the lower limit of the re-calculated sample size. Within the unrestricted design, the re-calculated sample size is permitted to be lower than the initially planned one while within the restricted design; the re-calculated sample size can only be larger than the initially planned one. Less attention has been paid to a possible upper limit of the sample size. The upper limit might either be given by the form of the sample size equation itself or by external factors like budget or prevalence of the disease. We will refer to this design as the double-restricted design meaning that there is a lower as well as an upper boundary of the re-calculated sample size.
We investigated the properties of the double-restricted design with respect to type I error, power, and expected sample size for different test statistics as well as different endpoints. It can be shown that the choice of the upper limit can dramatically affect the power of the resulting trial.

Adaptive Dose-selection in Equivalence Trials

Benjamin Lang[1], Cornelia Ursula Kunz

[1])Boehringer Ingelheim Pharma GmbH & Co. KG
benjamin.lang@boehringer-ingelheim.com

Drug development is very expensive and risky with many compounds failing in late development phases. Adaptive designs have been recognized as a way to improve efficiency of drug development by industry and regulators alike. Such designs use information from accumulating data in an ongoing trial to make decisions about the conduct of the rest of the study. Of particular interest are designs combining aspects of the clinical development process into one single study that would have traditionally been assessed in separate trials and phases, for instance, adaptive seamless phase II/III designs. A trial of this type is conducted in two stages: during the first stage, the exploratory stage, patients are recruited to several experimental treatments and a control treatment. During the second stage, the confirmatory stage, patients are enrolled to the remaining treatment arms or the control arm. The final analysis is based on the data from both stages. So far methodological research has focused on dose-selection for superiority trials. However, situations exist where the aim is to choose the treatment arm that is most similar to the control arm. We propose a method for adaptive dose-selection within an equivalence trial based on normally distributed endpoints. Analytical solutions are derived allowing determination of critical boundaries to control the type I error as well as the sample size for a given desired power. The proposed method also allows implementation of stopping for futility at interim. We illustrate the method using a recent trial example.

Optimal test procedures for multiple hypotheses controlling the familywise expected loss

Willi Maurer, Frank Bretz[1] and Xiaolei Xun

[1]) Novartis Pharma AG
frank.bretz@novartis.com

We consider the problem of testing multiple null hypotheses where a decision to reject or retain is to be made for each individual hypothesis. Based on the decision-theoretic framework, we propose to control the familywise expected loss instead of the conventional familywise error rate (FWER). Various loss functions can be adopted and the FWER is seen to result as a particular choice of the loss function. We search for decision rules that satisfy certain optimality criteria within a broad class of rules for which the expected loss is bounded by a pre-specified threshold under any parameter configuration. This approach is different from the canonical decision theory of maximizing a single utility function, but in analogy to classical hypothesis testing. We illustrate the methods with the problem of establishing efficacy of a new medicinal treatment in non-overlapping subgroups of patients.

Moderating the trade-off between type I and type II errors via the scaled false discovery rate

Djalel-Eddine Meskaldji[1], Stephan Morgenthaler

[1]) Applied statistics, Institute of mathematics
arthur.author@correspondence.email.com
[2]) Second affiliation

When many null hypotheses are tested, the control of the type I error is often the principal consideration. The scaled false discovery rate sFDR=E(V/s(R)), where V is the number of false positives, R is the number of rejections and s is a non-decreasing function, tunes the influence of the number of rejections in the control of type I errors thanks to the scaling function s. The sFDR control can be achieved via a step-up procedure with threshold sequence proportional to s, that is, it has s as a shape function. The sFDR gives the ability of moderating the trade-off between type I and type II errors with one well chosen threshold sequence anticipating diverse scenarios involving weak or strong effects. With elements of an optimality theory, by considering the number of false rejections V separately from the number of correct rejections S, we discuss the flexibility offered by the sFDR and the problem of how to choose which error rate and which procedure to use in practice.

An adaptive design for a self-limiting disease

Bart Michiels[1], Wilbert van Duijnhoven

[1]) Johnson & Johnson, Janssen R&D
bmichiel@its.jnj.com

To evaluate the efficacy of a new experimental treatment, to be used in patients infected with a self-limiting seasonal disease to speed up their recovery, a phase 3 study was designed.
Due to the high degree of uncertainty on the primary endpoint, an adaptive design has been incorporated allowing to stop the study early in case of lack of effect (futility), and to increase the study sample size (sample size re-estimation). The interim evaluation, at a pre-specified timepoint taking into account the seasonality of the disease, will be based on the available unblinded data and handled through an Independent Data Monitoring Committee (IDMC).
The primary endpoint in the study is time to resolution of symptoms and is assumed to follow a log-logistic distribution, with the treatment effect captured through an accelerated failure time (AFT) model.
Because of lack of an analytical approximation for an AFT, the corresponding sample size calculations were done through simulations – posing additional complexities to implement the adaptations. Instead of requiring that the IDMC would run extensive simulations to obtain the updated sample size during the interim analysis, an approximate formula was derived which could be used to adjust the sample size. The accuracy of the formula is shown through simulations.
Once the new sample size is set, the IDMC will evaluate the futility of the study using a conditional power approach.
The characteristics of the adaptive design were evaluated through extensive simulations under various conditions.

Adaptive MCPMod Testing

Tobias Mielke[1], Franz König

[1] Janssen, QS Consulting
tmielke1@its.jnj.com

The MCPMod procedure for model-based dose-finding under model uncertainty has received considerable attention in the scientific literature during the last years. MCPMod combines multiple model-based trend tests with dose-response modelling approaches. Adaptive applications of MCPMod may include adaptations to the randomization ratios for the doses included into a study, as well as the set of contrast coefficients used to test for existence of drug related effects. Appropriate statistical methods need to be applied to ensure a strong error control in case of adaptations to the set of contrast coefficients based on unblinded data. The inverse normal combination approach may be straight forwardly generalized from seamless Phase 2/3 designs with multiple treatment groups to the situation of adaptive MCPMod testing using multiple contrast vectors. However, while the inverse-normal method is generally efficient in case that only one test arm is forwarded into the following study stages, it may lose efficiency in case that multiple test arms are forwarded. In adaptive applications of the MCPMod approach, it could happen in particular that all contrast vectors are forwarded, such that an inverse normal combination approach might result in reduced efficiency. To avoid such inefficiency the conditional error principle [1] as suggested for the adaptive Dunnett test can be applied to address design modifications when testing several contrasts. We will generalize the conditional error principle to the situation of adaptive MCPMod testing. We discuss potential benefits and limitations as compared to the inverse-normal combination approach.

References

[1] Müller, H.H., & Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. Statisics in Medicine 23, 2497-2508

Blinded continuous information monitoring of recurrent events endpoints with time trends

Tobias Mütze[1], Salem, Susanna; Benda, Norbert; Schmidli, Heinz; Friede, Tim

[1]) Novartis Pharma AG
tobias.muetze@novartis.com

In clinical trials with recurrent event endpoints, misspecified assumptions of event rates or the dispersion can lead to under- or overpowered trials. Specification of the overdispersion is often a particular problem as it is usually not reported in clinical trial publications. To mitigate the risks of inadequate sample sizes, internal pilot study designs for clinical trials with recurrent events have been proposed, with a preference for blinded sample size re-estimation procedures as they generally do not affect the type I error rate and maintain trial integrity [1]. However, the re-estimated sample size can have considerable variance, in particular with early sample size reviews. Friede et al. (2018) [2] addressed the issue of variable re-estimated sample sizes by proposing a blinded continuous monitoring of information for clinical trials with recurrent events modelled by a homogeneous Poisson process with a Gamma frailty. However, the assumption of a time-independent event rate in a homogeneous Poisson process does not always hold. For example, Nicholas et al. (2012) [3] showed that the relapse rate in clinical trials in multiple sclerosis changes over time. In this presentation, we study the robustness of the continuous information monitoring procedure proposed by Friede et al. (2018) [2] towards recurrent events with time trends. Moreover, we propose a blinded continuous information monitoring procedure for recurrent events with time trends. We show that our proposed monitoring procedure does maintain the integrity a clinical trial by controlling the type I error rate and that the proposed procedure results in adequately powered clinical trials.

References

[1] Friede, T., Schmidli, H. (2010). Blinded sample size reestimation with negative binomial counts in superiority and non-inferiority trials. Methods of Information in Medicine, 49:618-624.

[2] Friede, T., Häring, D., Schmidli, H. (2018). Blinded continuous monitoring in clinical trials with recurrent event endpoints. (Submitted).

[3] Nicholas R., Straube S., Schmidli H., Pfeiffer S., Friede T. (2012). Time-patterns of annualized relapse rates in randomized placebo-controlled clinical trials in relapsing multiple sclerosis: a systematic review and meta-analysis. Multiple Sclerosis, 18:1290-1296.

Estimating the proportion of true null hypotheses under arbitrary dependency

Andre Neumann[1], Taras Bodnar, Thorsten Dickhaus

[1])University Bremen, Institute for Statistics
neumann@uni-bremen.de

It is a well known result in multiple hypothesis testing that the proportion π_0 of true null hypotheses is not identified under general dependencies. However, it is possible to estimate π_0 if structural information about the dependency structure among the test statistics or p-values, respectively, is available. In this talk I demonstrate these points, and explain our proposed marginal parametric bootstrap method. A pseudo-sample of bootstrap p-values is generated, which still carry information about π_0, but behave like realizations of stochastically independent random variables.

Oncology Phase II Adaptive Designs - Treatment effect estimates and their use in planning Phase III trials

Arsénio Nhacolo[1], Werner Brannath

[1]) University of Bremen, KKSB
anhacolo@uni-bremen.de

New estimation methods for oncology Phase II adaptive designs

We propose point and interval estimation for adaptive designs. We considered the recently proposed oncology Phase II two-stage single-arm adaptive designs with binary endpoint, in which the second stage sample size is a predefined function of the first stage's number of responses. Our approach is based on sample space orderings, from which we derive p-values, and point and interval estimates. Simulation studies show that our proposed methods perform better, in terms of bias and root mean square error, than the fixed-sample maximum likelihood estimator.

Using Estimates from adaptive Phase II oncology trials to plan Phase III trials

The clinical drug development is mainly done in three phases, Phase I, Phase II and Phase III. The knowledge gained in clinical trials of a particular phase is often used to plan trials of subsequent phases. That is the case with successful Phase II clinical trials in which, among others aspects, the effect size estimates are used to plan the sample size of the related Phase III trials. Due to small sample sizes, selections bias and other factors, Phase II estimates are often imprecise, resulting in inadequately powered Phase III trials. We evaluated through simulation studies the consequences, in terms of power, of using the effect estimate from Phase II adaptive design trials to plan sample size of Phase III trials in oncology. In addition, we propose a new approach for adjusting Phase II estimates. We used the naïve maximum likelihood and our proposed estimators for estimating the Phase II effect. Results showed that using naïve estimates lead to underpowered Phase III trials, while estimates that take into account the adaptiveness of the designs lead to power that is close to the target value. Our new adjustment approach seems to perform well for all estimation methods. It also showed that a relatively higher discount is necessary for naïve estimates.

Optimal adaptive sample size recalculation for normally distributed outcomes

Maximilian Pilz[1], Kevin Kunzmann, Carolin Herrmann, Geraldine Rauch, Meinhard Kieser

[1]) University of Heidelberg, Institute of Medical Biometry
pilz@imbi.uni-heidelberg.de

In clinical trials, the choice of an adequate sample size is a crucial issue. While traditionally clinical trials were performed with fixed sample size, application of designs with the option of interim sample size recalculations becomes increasingly popular. Adaptive two-stage designs allow a sample size recalculation after a planned unblinded interim analysis in order to adjust the sample size during the ongoing trial. Various adaptive approaches exist differing, e.g., by decision boundaries, sample size recalculation rule, and first-stage sample size.

In the planning phase, one is faced with the challenge to choose the most appropriate design for the present study. However, comparison of the various methods is not straightforward. There are many possibilities how to choose a criterion that evaluates the performance of a design. We focus on expected sample size of the trial under the alternative hypothesis to compare different approaches.

When the performance criterion is fixed, it is natural to choose a design optimizing it. Jennison and Turnbull (2015) were the first analyzing optimal adaptive designs. We extend their approach and obtain a design which outperforms the design by Jennison and Turnbull. Both analytical and numerical methods were developed in order to derive the design that optimizes the chosen performance criterion while at the same time fulfilling important constraints as type I error and power restrictions. These considerations are made without restrictive assumptions on the design parameters as, e.g., a combination test. Therefore, the resulting designs are optimal under a wide class of designs. A comparison to other designs as, e.g., the design by Jennison and Turnbull or classical designs based on combination functions with sample size recalculation using conditional power demonstrates the differences.

Mastering Multiplicity in Clinical Trials: Shortcuts, Graphs and Error Rates

Martin Posch[1]

[1]) Medical University of Vienna
martin.posch@meduniwien.ac.at

Multiplicity is a fundamental issue for statistical inference in clinical trials whenever multiple endpoints, subgroups or treatment arms are considered or multiple analysis are performed, as e.g., in group sequential trials. Especially in the confirmatory setting, regulatory guidelines require to account for multiplicity to control the family wise error rate. In this talk I highlight some of Willi Maurer‘s fundamental contributions to the theory of multiple testing, addressing his work on the construction of shortcuts for closed, consonant Bonferroni type tests and the theory of graphical multiple testing procedures. Furthermore I touch on his seminal work going beyond the control of the FWER, towards control of an expected loss. The latter concept is especially relevant for the analysis of platform trials, a new type of trials where sub-studies testing separate hypotheses may be added or dropped over time.

A new semi-Bayesian rule for sample size recalculation in adaptive designs

Geraldine Rauch[1], Meinhard Kieser, Maximilian Pilz, Kevin Kunzmann, Carolin Herrmann

[1]) Charité – Universitätsmedizin Berlin and Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology
geraldine.rauch@charite.de

It is intuitive that the correct choice of the sample size is of major importance for an ethical justification of a trial and a responsible spending of resources. In an underpowered trial, the research hypothesis is unlikely to be proven, resources are wasted and patients are unnecessarily exposed to the study-specific risks. If the sample size is too large, the market approval is prolonged and later recruited patient in the control arm are exposed to a treatment already known to be less effective. The parameter assumptions required for sample size calculation should be based on previously published results from the literature and on aspects of clinical relevance. In clinical practice, however, historical studies for the research topic of interest are often not directly comparable to the current situation under investigation or simply do not exist. Moreover, the results of previous studies often show a high variability or are even contradictory.
Calculating the 'correct' sample size is thus a difficult task. On the other side, the consequences of a 'wrong' sample size are severe. A variety of sample size recalculation strategies has been proposed. Most frequently, these rules are based on conditional power arguments (e.g. [1], [2], [3]). This approach assumes implicitly that the true treatment effect is equal to the effect observed at the interim analysis. The conditional power approach is often criticized for this unrealistic assumption as the available information at the interim stage is usually limited and thus the treatment effect estimate shows a rather high variability resulting in a highly variable sample size. Built upon the new insights we gained from developing a new performance score for adaptive designs (presented in the related talk 'A new conditional performance score for evaluating sample size recalculation rules in adaptive designs'), we present a new semi-Bayesian sample size recalculation strategy which uses the interim effect as the expectation of a prior distribution rather than assuming that the interim effect is the true one.

References

[1] Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. Biometrics 1999, 1286-1290.

[2] Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: A practical guide with examples. Stat. Med. 2011, 30: 3267-3284.

[3] Jennison C, Turnbull BW. Adaptive sample size modification in clinical trials: start small then ask for more? Stat. Med. 2015, 34: 3793–3810.

Online control of the false discovery rate in biomedical research

David Robertson[1], James Wason

[1])University of Cambridge, MRC Biostatistics Unit
david.robertson@mrc-bsu.cam.ac.uk

Modern biomedical research frequently involves testing multiple related hypotheses, while maintaining control over a suitable error rate. In many applications the false discovery rate (FDR), which is the expected proportion of false positives among the rejected hypotheses, has become the standard error criterion. Procedures that control the FDR, such as the well-known Benjamini-Hochberg procedure, assume that all p-values are available to be tested at a single time point. However, this ignores the sequential nature of many biomedical experiments, where a sequence of hypotheses is tested without having access to future p-values or even the number of hypotheses (potentially infinite) to be tested. Recently, procedures that control the FDR in this online manner have been proposed by Javanmard and Montanari (Ann. Stat. 46:526-554, 2018), and built upon by Ramdas et al. (arXiv 1710.00499, 1802.09098). In this talk, we compare and contrast these proposed procedures, with a particular focus on settings where the p-values are dependent and where the number of hypotheses to be tested is not very large. We also propose a simple modification of the procedures for when there is an upper bound on the number of hypotheses to be tested. Using comprehensive simulation scenarios and case studies, we provide recommendations for which procedures to use in practice for online FDR control.

More efficient treatment effect estimation in pre-specified subgroups displayed in forest plots for time-to-event outcomes

Kaspar Rufibach[1], Marcel Wolbers, Ke Li

[1]) F. Hoffmann-La Roche, Methods, Collaboration, and Outreach Group, Department of Biostatistics
kaspar.rufibach@roche.com

In randomized controlled trials, the homogeneity of treatment effect estimates in pre-defined subgroups based on clinical, laboratory, genetic, or other baseline markers is frequently investigated using forest plots. However, the interpretation of naïve subgroup-specific treatment effect estimates requires great care because of the smaller sample size of subgroups (implying large variability of estimated effect sizes) and the frequently large number of investigated subgroups. Treatment effect estimates in subgroups with a lower mean-square error based on frequentist and Bayesian shrinkage, Bayesian model averaging, and the bootstrap have recently been investigated but focused on continuous outcomes. We propose two novel general strategies for treatment effect estimation in subgroups for survival outcomes. The first strategy is to build a flexible model based on all available observations including all relevant subgroups and subgroup-treatment interactions as covariates. This model is then marginalized to obtain subgroup-specific effect estimates. We propose to use the average hazard ratio corresponding to the odds of concordance for this marginalization. The second strategy is based on simple subgroup-specific models which are combined via (penalized) composite likelihood. We implement these strategies to obtain shrinkage estimators using lasso and ridge penalties. With this, we can interpolate between the two extreme scenarios of either taking the overall estimate as best estimate in every subgroup, or computing effect estimates within each subgroup separately. We illustrate under which scenarios this strategy provides a pronounced improvement in mean squared error compared to the extreme strategies. The methods are illustrated with data from a large randomized registration trial in follicular lymphoma.

Multiple testing approaches in dose combination trial

Saswati Saha[1], Werner Brannath, Bjoern Bornkamp

[1]) University of Bremen, Applied Statistics and Biometry
saha@uni-bremen.de

Drug combination trials are often motivated from the fact that individual drugs target the same disease but via different routes. So combining drugs to obtain an overall better effect is more efficient than conducting individual treatment. Often we come across diseases such as cancer and severe asthma on which the standalone drug does not yield expected results. It is to cater to these instances that combining drugs becomes a necessity sometimes. Several approaches have been explored for developing statistical methods that compare (single) fixed dose combination therapies to its component. But extension of these approaches to the situation where multiple dose combinations are compared against their components is not always easy. We propose two approaches by which one can provide confirmatory assurance with overall control of type 1 error, that combination of two drugs is more effective than either component drug alone. These approaches involved multiple comparisons in multilevel factorial design where the overall type 1 error can be controlled firstly, by bootstrap test, and secondly, by considering the least favorable null configurations under a union intersection test. In this presentation we would like to demonstrate the implementation of these new approaches with a real data example from a blood pressure reduction trial and via extensive simulations show how the new approaches perform when bench marked with an existing approach.

Analysis strategies for adaptive survival trials with multiple time-to-event endpoints

Rene Schmidt[1], Prof. Dr. Andreas Faldum

[1]) WWU Münster, Institut für Biometrie und Klinische Forschung (IBKF)
rene.schmidt@ukmuenster.de

The theory of adaptive designs is now well understood for short-term endpoints when the outcome of the patients is observed more-or-less immediately. For such settings, adaptive designs were first proposed by Bauer (1989) and Bauer and Köhne (1994) and reach their full potential in the work of Brannath et al. (2002) and Hommel (2001). For survival endpoints, however, subtle problems arise. Statistical challenge in adaptive survival trials is dealing with patients who enter the trial prior to an interim analysis and remain event-free beyond the interim analysis (c.f. Bauer and Posch, 2004). Historically first adaptive survival tests were constructed using the independent increments property of the log-rank statistic. These methods essentially only work if interim decision making is based solely on the interim log-rank statistic. Simultaneous use of data from several survival endpoints for design modifications is in general not admissible. Alternative approaches (based on the "patient-wise separation" principle) allow design modifications to be based on the full interim data, however, with the common disadvantage that the final test statistic may ignore part of the observed survival times or that the critical boundaries have to be adjusted, thus resulting in a conservative test procedure. In this talk, we discuss these current analysis strategies for adaptive survival trials and propose some new ideas how to design adaptive survival trials.

References

[1] Bauer, P. (1989). Multistage testing with adaptive design. Biometrie und Informatik in Medizin und Biologie 21: 1043 – 1066.

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

[3] Bauer, P. and Posch, M. (2004). Letter to the editor: modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. Statistics in Medicine 23: 1333 – 1335.

[4] Brannath, W. and Posch, M. and Bauer, P. (2002). Recursive combination tests. Journal of the American Statistical Association 97: 236 – 244.

[5] Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. Biometrical Journal 43: 581 – 589.

Testing endpoints with unknown correlation

Susanne Urach[1], Franz König, Martin Posch

[1]) Medical University of Vienna, Section for Medical Statistics, CEMSIIS
susanne.urach@meduniwien.ac.at

As the correlation structure is usually unknown, multiple testing procedures of endpoints in confirmatory clinical trials often use conservative methods based on the marginal distributions of test statistics to strongly control the family-wise type I error rate. Alpha exhaustive tests relying on the joint multivariate distribution either presume known correlations or use estimates based on the observed data. Calculating the critical boundaries under the assumption that the correlations are equal to some known values can lead to a type I error rate inflation in case of misspecification, the same is true if the correlations are estimated from the data and the sample size is low. We considered multiple testing procedures where the critical values are derived based on the assumption of multivariate normally distributed test statistics and quantified the inflation of the type I error rate due to assumed and estimated correlations. Furthermore, we apply the confidence interval approach by Berger and Boos to the two endpoint setting in order to deal with the unknown correlation and achieve strict type I error rate control for bivariate normally distributed test statistics. We improved Berger and Boos' method by deriving a sharper upper bound for the type I error rate increasing the power of the corresponding multiple testing procedure. The impact of using t-distributed instead of normally distributed test statistics is evaluated and respective adjustments are explored. The power of the methods assuming known correlation respectively estimating the correlation and the Berger Boos method are compared with non-parametric testing procedures in the setting with two endpoints.

Improving interim decisions in randomized trials by exploiting information on short-term outcomes and prognostic baseline covariates

Kelly Van Lancker[1], An Vandebosch and Stijn Vansteelandt

[1])Ghent University, Department of Applied Mathematics, Computer Science and Statistics
kelly.vanlancker@ugent.be

Interim analyses are routinely used to monitor accumulating data in clinical trials. A problem in such analyses is that all patients may have been enrolled by the time a sufficient number of patients have their primary endpoint available. When the objective of the interim analysis is to stop the trial when treatment is futile, it must ideally be conducted prior to enrollment completion. To remedy this problem, we propose an interim decision procedure which exploits the information contained in baseline covariates and short-term outcomes that are predictive of the final outcome. We show that the proposed procedure leads to a gain in efficiency, an increased power and a reduced sample size, without compromising the Type I error rate of the procedure, even when the used prediction models are misspecified. In particular, implementing our proposal in the conditional power approach allows earlier stopping for true futility whilst controlling the probability for incorrectly stopping. This has the consequence of reducing the number of recruited patients in case of stopping for futility, such that fewer patients get the futile regimen. In addition, we extend the method to adaptive designs with unblinded sample size reassessment based on conditional power arguments using the inverse normal method as the combination function. We support the proposal by simulation studies based on data from a real clinical trial.

Estimation of Sample Size and Power for Dunnett's Testing Setups with Unequal Effect Sizes

Marcus Vollmer[1]

[1])Institute of Bioinformatics, Universtity Medicine Greifswald
marcus.vollmer@uni-greifswald.de

Background: Dunnett's T3 procedure is a standard statistical test when comparing multiple treatment groups with the same reference group. Especially in animal experiments it is common to compare for example different immunized mice with an unimmunized control group. Interestingly, in animal experiments equal sample sizes have been frequently proposed. However, the same statistical power can be achieved by unequally distributed group sizes with a reduction in the total sample size.

Currently, two packages are available in R (multcomp, DTK) to perform the special testing problem with unequal group sizes. The computation of the p-values includes the consideration of a multidimensional t-distribution and the adjustment for multiple testing. R:DunnettTests conducts a sample size calculation, but only with identical treatment effect size and pre-specified sample allocation ratio.

Methods: We developed a method to derive unequal group sizes while assuming different effect sizes (different means and unequal variances). The method minimizes the number of animals needed in such experiments while performing Dunnett's T3 procedure. The minimal set of group sizes was derived using a genetic algorithm on Monte Carlo experiments. A topological concept on integer partitions was used for finding the optimal set of group sizes in a timely manner.

Results: For different effect and group sizes, the total sample size reduction ranges between 5 % and 20 % through imbalanced testing, which has a directly impact on experimental costs. The higher the number of treatment groups the higher is the sample size reduction compared to a balanced layout. The topological concept of searching for an optimal set of sample sizes can be easily transferred to other testing problems. A case example for sample size justification for an animal test proposal is given.

RPACT An R Program for Confirmatory Adaptive Group Sequential Designs

Gernot Wassmer[1] and Friedrich Pahlke

[1])University of Cologne, RPACT GbR
gernot.wassmer@uni-koeln.de

There is increasing interest by the industry to use R. At the moment, no R package is available for performing confirmatory adaptive designs in a comprehensive sense (e.g., design and analysis for continuous, binary, and survival endpoint). Nevertheless, for group sequential tests there is the R package gsDesign, developed by Keaven Anderson (copyright Merck Research Laboratories), which is well established and covers many relevant designs. Among the over 12.000 available packages at CRAN (July 2018) there are several packages that address the issue of adaptive designs, most of them with special reference to research results from the authors, but none covers the broad range of applications that is nowadays available. In RPACT (R Package for Adaptive Clinical Trials) particularly, the methods described in the recent monograph of Wassmer and Brannath (published by Springer, 2016) are implemented and made available for the public.

We describe the basic features of the current version of RPACT. For design and analysis, this includes all relevant cases for group sequential designs without sample size re-estimation, adaptive designs that are based on the inverse normal method, and adaptive designs that are based on Fisher's combination test. For analysing the data, besides assessing conditional properties (i.e., conditional power and conditional rejection probability (CRP) under H0) confidence intervals and p-values that account for the adaptive nature of the designs are provided. The validation of the package will be done compliant to FDA/GxP guidelines and to the validation process of "Base R" and "Recommended Packages" as described in: "R: Regulatory Compliance and Validation Issues, A Guidance Document for the Use of R in Regulated Clinical Trial Environments" (The R Foundation for Statistical Computing, December, 2014).

Pre-planned subgroup analysis within a group-sequential design for a time-to-event endpoint

Marcel Wolbers[1], Kaspar Rufibach

[1])F. Hoffmann-La Roche AG, Methods, Collaboration, and Outreach Group;
Biostatistics Department
marcel.wolbers@roche.com

For targeted chemotherapies it is often unclear whether the treatment benefit is restricted to a biomarker-positive subgroup or applies to the full trial population. We present group-sequential designs for the joint-evaluation of a time-to-event endpoint in two co-primary populations: the full population and a pre-defined subgroup of it. The proposed group-sequential boundaries allow to pre-assign importance weights to the two co-primary populations and exploit correlations of test statistics across interim analyses and between data from the subgroup and the full population. The basic idea is to use an alpha-spending function to determine how much type I error can be spent at each interim analysis across both populations and to subsequently determine corresponding critical values for each sub-population separately. In a case study, the method will be compared to alternative development approaches demonstrating its advantage over purely hierarchical strategies which designates either of the populations as primary and the other as secondary. Extensions to more than two populations will also be discussed.

A new omnibus test for the global null hypothesis

Sonja Zehetmayer[1]

[1])Medical University of Vienna, Center for Medical Statistics, Informatics, and Intelligent Systems
sonja.zehetmayer@meduniwien.ac.at

Global hypothesis tests are an important tool in the context of, e.g, clinical trials, genetic studies or meta analyses, when researches are not interested in testing individual hypotheses, but in testing whether none of the hypotheses is false. There are several possibilities how to test the global null hypothesis when the individual null hypotheses are independent. If it is assumed that many of the individual null hypotheses are false, combinations tests (e.g, Fisher or Stouffer test), which combine data from several endpoints to a single test statistic, have been recommended to maximise power. If, however, it is assumed that only one or a few null hypotheses are false, global tests based on individual test statistics are more powerful (e.g., Bonferroni or Simes test). However, usually there is no a-priori knowledge on the number of false individual null hypotheses. We therefore propose an omnibus test based on the combination of p-values. We show that this test yields an impressive overall performance.

ABSTRACTS (POSTER PRESENTATIONS)

Optimising the sample allocation across a multi-stage adaptive confirmatory clinical trial

Nicolas Ballarini[1], Thomas Burnett, Thomas Jaki, Christopher Jennison, Franz König, Martin Posch

[1]) Medical University of Vienna
nicolas.ballarini@meduniwien.ac.at

We design multi-stage adaptive confirmatory clinical trials that make use of Bayesian decision theoretic framework and a utility function which may take into account the prevalence of the subpopulations, costs, or the true treatment effects. Given a pre-specified utility function, our proposal allows altering the sample allocation and hypothesis testing weights at any stage of the trial, ensuring efficient use of available resources to maximize the expected utility. This design englobes Adaptive Enrichment and single-stage designs as special cases. We consider testing the elementary null hypotheses of disjoint subgroups and guarantee strong control of the Familywise Error Rate using the conditional error rate approach.

We present the results of simulation studies in a variety of cases to compare the effectiveness of the optimal sample allocation design with Adaptive Enrichment and fixed sampling alternatives.

The optimisation is not necessarily restricted to optimisation within a single trial. We discuss extensions that consider optimal budget allocation for simultaneous adaptive clinical trials.

On a New Two-Sample Log-Rank Test

Laura Kerschke[1], Andreas Faldum, Rene Schmidt

[1]) University of Münster, Institute of Biostatistics and Clinical Research
laura.kerschke@ukmuenster.de

The two-sample log-rank test proposed by Mantel [1] and Peto and Peto [2] is the most commonly used test when survival curves of two treatment groups are compared within a clinical trial. It can easily be generalized to handle more than two groups or to adjust for covariate effects using a proportional hazards model. As shown by Peto and Peto [2], the log-rank test is optimal under the proportional hazards condition and even remains valid in the case of non-proportional hazards. Despite these favorable properties, methodological difficulties arise in adapting the log-rank test to more complex study designs as e.g. platform trials that allow to evaluate several treatments across one or more types of patients in a single trial, thus increasing efficiency of drug development process. This is due to the fact that the common two-sample log-rank test statistic is obtained from pooled data of both treatment groups and cannot be written as the difference of two independent random variables derived from non-overlapping populations. To overcome this issue, we propose an alternative two-sample log-rank test such that the underlying test statistic equals the difference of two independent random variables derived from non-overlapping populations. The test statistic is thus similar to that of an unpaired t-test with known variance. On this basis the well-known methodology for comparisons of means can immediately be transferred to the survival setting. The proposed method relies on asymptotic distributions. We study its performance in the two-sample single-stage setting. Our simulations support validity of the distributional approximations as well as adequate type I and type II error rate control. Elaboration of specific platform trial designs will be content of further research.

References

[1] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports. 50, 163-170.

[2] Peto, R. and Peto J. (1972). Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society, Series A. 135, 185-207.

Homogeneity analyses in frequentist approaches of evaluation of basket trials.

Maja Krajewska[1], Geraldine Rauch

[1]) Charité Universitätsmedizin Berlin, Institute of Biometry and Clinical Epidemiology
maja.krajewska@charite.de

Background: Traditionally, treatments in oncology are mainly determined based on the anatomical location of the tumor, but recent findings suggest also taking into account the genetic mutations of the tumor. In accordance with these findings, the effectiveness of a treatment on tumors with the same genetic mutations can be analyzed in so-called basket trials. The aim of these trials is to examine the effectiveness of a treatment in patients with different types of a disease but exhibiting the same biomarkers. The application of this design in the oncological setting is based on the accrual of patients exhibiting tumors in different anatomic locations but the same genetic mutation. After accrual, patients are allocated into subgroups, so-called "baskets", based on the anatomical location of their tumor. Propositions for the evaluation of these trials have been mainly based on hierarchical Bayesian modeling [1, 2, 3], which allows for the exchange of information among baskets judged as homogeneous. These approaches are computationally complex and require prior assumptions about the efficacy of the treatment, which might be difficult to make in early phase trials. Therefore, Cunanan et al. [4] proposed a frequentist approach based on the concept of homogeneity of baskets: a design in which all baskets are tested for homogeneity at the interim analysis and subsequently either pooled or evaluated individually. The decision is made based on Fisher's exact test, which test the independence of baskets using a homogeneity parameter that is set beforehand.
Objective: The objective of this work is to examine the performance of the homogeneity analysis in the design proposed by Cunanan et al. [4], with an emphasis on the frequency of correctly assigned baskets by the homogeneity analysis as well as on the discussion of alternative design options.
Proposed Methods: In order to examine the efficiency of the homogeneity analysis, we recreated the proposed algorithm in R [5]. For the simulation, we used the parameters reported by Cunanan et al. [4]: number of baskets $K = 5$ and sample size per basket $n = 7$; the treatment was judged ineffective if the tumor response rate was smaller than $theta_0 = 0.15$ and as effective if it was bigger than or equal to $\theta_a = 0.45$. The simulations were performed for all six scenarios reported by Cunanan et al. [4]: the treatment is effective in none, in one, in two, in three, in four or in all baskets. All scenarios were simulated for both equal and varying accrual of patients into the individual baskets. The number of correctly assigned baskets by the homogeneity analysis was calculated as well as the number of false positive and false negative results of baskets wrongly judged homogenous.
Discussion: Based on the results of the simulations, the efficiency of the judgment of baskets by the homogeneity analysis will be discussed, as well as the

consequences of incorrect pooling of baskets. Additionally, we discuss whether Fisher's exact test represents the best choice for judging the homogeneity of baskets. As an alternative to the design proposed by Cunanan et al. [4], we discuss possible designs based on judging all baskets as heterogeneous and on pooling individual baskets.

Practical considerations of MCP-Mod in applying to multi-regional dose finding studies

Toshifumi Sugitani[1], Yusuke Yamaguchi

[1]) Astellas Pharma Inc., Biostatistics Group
toshifumi.sugitani@astellas.com

In this talk, we share our practical experience in applying MCP-Mod approach to a multi-regional dose finding study, where we face some inherent issues to consider. For example, in such a study, we are better off combining the MCP-Mod with hypothesis testing of efficacy of doses, since it is often the case that the Japanese regulatory agency (PMDA) thinks of dose-finding study as a part of confirmatory study. Other examples include assessment of dose-response similarity between regions (e.g, Japan vs overall population), and consideration of sample size allocation ratio to be able to have consistent results between individual region(s) and overall population, which is usually required by PMDA. In the talk, we present simulation results to discuss these topics.

Evaluation of multiple prediction models

Max Westphal[1], Werner Brannath

[1]) University of Bremen, Institute for Statistics
mwestphal@uni-bremen.de

Model selection and performance assessment for prediction models are important and difficult tasks in machine learning. A common approach is to select a single model via cross-validation and to evaluate this final model on an independent dataset. In this scenario, it is usually not difficult to conduct statistical inference on the generalization performance of the chosen model.
We propose to instead evaluate several models simultaneously. These may result from varied hyperparameters or completely different learning algorithms. Our main goal is to increase the probability to correctly identify a model that performs sufficiently well. In this case, adjusting for multiplicity is necessary in the evaluation stage to avoid an inflation of the family wise error rate.
We apply the so-called maxT-approach which is based on the distribution of the maximum test statistic and show that this approach is suitable to (approximately) control the family wise error rate for a wide variety of performance measures. In our framework, the final model selection is conducted on the evaluation data.
This strategy proved to be beneficial in simulation studies regarding statistical power and additionally the performance of the final model. Furthermore, we show how the introduced bias of performance estimates can be corrected.
We conclude that evaluating only a single final model is suboptimal. Instead, several promising models should be evaluated simultaneously to counter uncertainty in the (cross-)validation ranking. In particular, this strategy increases the probability to correctly identify a good model.

References

[1] Westphal, Max and Werner Brannath. Evaluation of multiple prediction models: A novel view on model selection and performance assessment. Manuscript submitted for publication. (2018).