daptive Designs and Multiple Testing Procedures

Workshop July 5 – 6, 2012 in Heidelberg

Conference Website:
www.biometrie.uni-heidelberg.de/workshop

Local Organizing Committee:
Meinhard Kieser, Kathrin Stucke, Stefan Englert, Alexander Kurz

Internationale Biometrische Gesellschaft Deutsche Region

IBS-ROeS

imbi Heidelberg Institute of Medical Biometry and Informatics

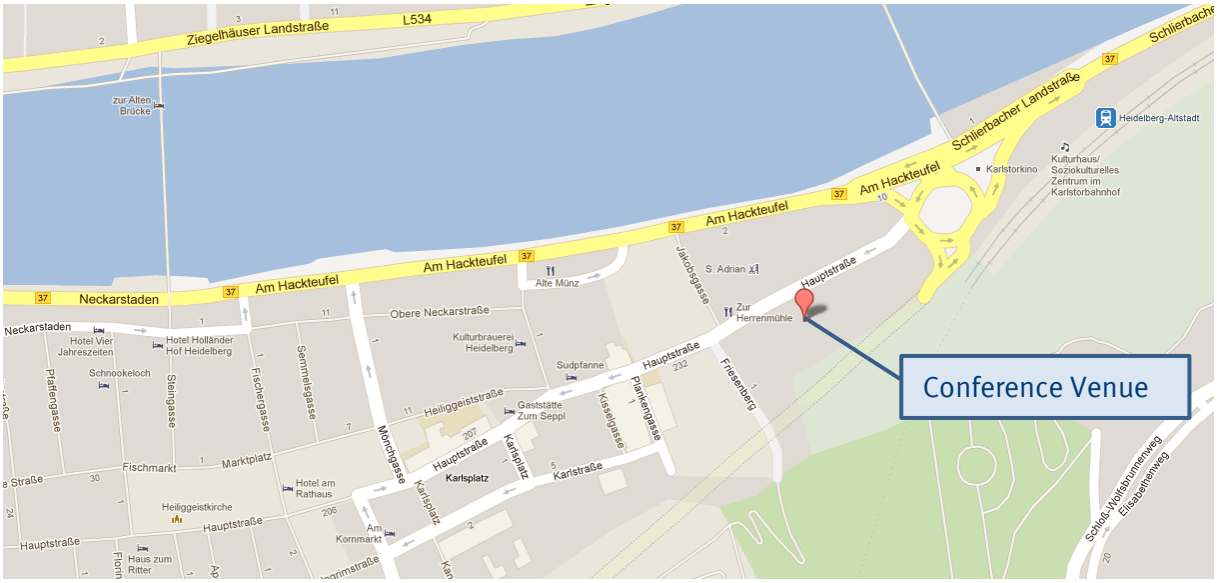
CONFERENCE VENUE

LOCATION:

The workshop "Adaptive Designs and Multiple Testing Procedures 2012" will take place at the Internationales Wissenschaftsforum Heidelberg (IWH). It is a centre sponsored by Heidelberg University for scholarly exchange in all areas of science and academic research.

ADDRESS:

Hauptstrasse 242
D-69117 Heidelberg
Tel: +49 (0) 6221 54 36 90
Fax: +49 (0) 6221 54 161 3691
E-mail: iwh@uni-hd.de

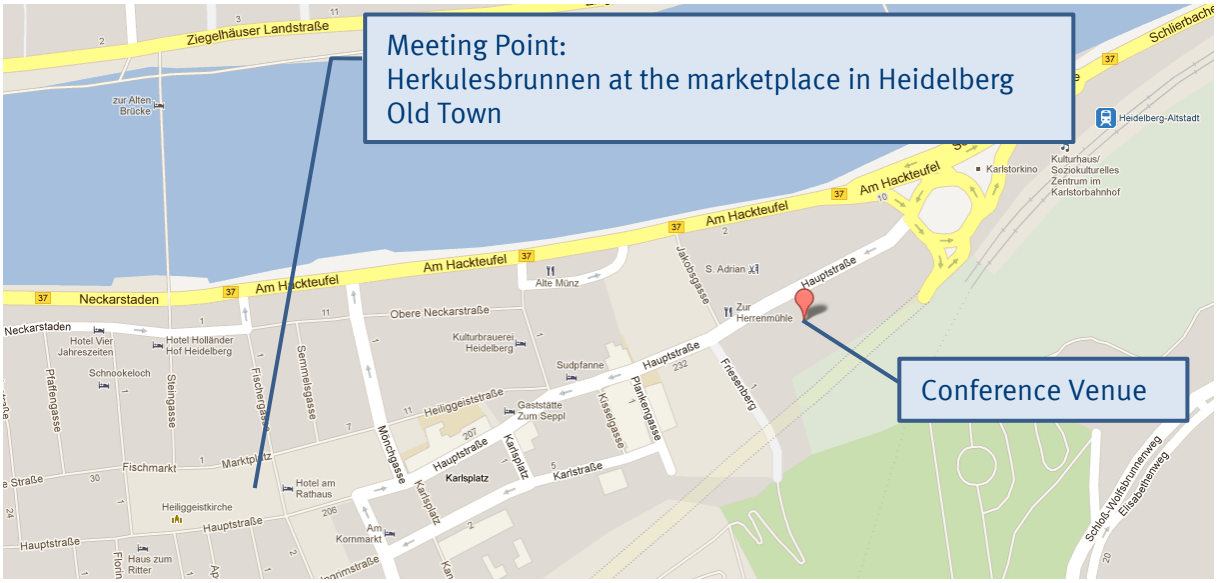


GUIDED TOUR

A guided tour can be attended on July 5 after the workshop sessions. The tour starts at 19:15 at the "Herkulesbrunnen" (Heidelberg Market Square).

HOW TO GET THERE:

The meeting point is in walking distance to the conference venue in the centre of Heidelberg Old Town.



Conference Dinner

The conference dinner can be attended on July 5 at 20:00 after the guided tour. The conference dinner will take place in the “Palmbräu Gasse”.

Address:

Palmbräu Gasse KERO GmbH
Hauptstraße 185
69117 Heidelberg

How to get there:

The location of the conference dinner is in walking distance to the conference venue. The guided tour will end directly at the location of the conference dinner.

SCIENTIFIC PROGRAM – OVERVIEW

THURSDAY, JULY 5

09:00 – 10:30	Registration and Reception
10:30 – 10:45	Welcome Addresses: Prof. Dr. Joachim Kirsch (Vice Dean, Medical Faculty Heidelberg) and Prof. Dr. Meinhard Kieser (Director Institute of Medical Biometry and Informatics, Heidelberg)
10:45 – 12:30	Session 1: Sample Size Re-Estimation
12:30 – 13:30	Lunch break
13:30 – 15:15	Session 2: Multiple Testing (1)
15:15 – 15:45	Coffee break
15:45 – 16:45	Invited Session
16:45 – 17:00	Coffee break
17:00 – 18:45	Session 3: Seamless Phase II/III Designs and Dose-Finding
19:15 – 20:00	Guided Tour
20:00	Conference Dinner

FRIDAY, JULY 6

08:30 – 10:15	Session 4: Time-to-Event Data and Confidence Intervals
10:15 – 10:30	Coffee break
10:30 – 12:15	Session 5: Multiple Testing (2)
12:15 – 12:30	Coffee break
12:30 – 14:15	Session 6: Selection of Treatments or Populations
14:15	Meeting of Working Group “Adaptive Designs and Multiple Testing Procedures” of IBS-DR and ROeS

SCIENTIFIC PROGRAM – DETAILED TIME SCHEDULE

THURSDAY, JULY 5

09:00 – 10:30 **REGISTRATION AND RECEPTION**

10:30 – 10:45 **WELCOME ADDRESSES:**
PROF. DR. JOACHIM KIRSCH (VICE DEAN, MEDICAL FACULTY HEIDELBERG)
PROF. DR. MEINHARD KIESER (DIRECTOR INSTITUTE OF MEDICAL BIOMETRY AND INFORMATICS, HEIDELBERG)

10:30 – 12:30 **SESSION 1: SAMPLE SIZE RE-ESTIMATION**
CHAIR: EKKEHARD GLIMM (BASEL), THOMAS JAKI (LANCASTER)

Simon Schneider (Göttingen), Heinz Schmidli (Basel): Blinded and unblinded internal pilot study designs for clinical trials with over-dispersed count data

Katharina Ingel, Antje Jahn-Eimermacher (Mainz): Adaptive sample size re-estimation for recurrent event data

Frank Miller (Södertälje/SWE), Tim Friede (Göttingen): Blinded continuous monitoring of the nuisance parameter in clinical trials

Stefan Englert, Meinhard Kieser (Heidelberg): Evaluation of sample size adaptation rules in clinical studies aiming at an overall performance optimization

Florian Klinglmüller, Franz König (Wien): Testing primary and secondary endpoints in adaptive designs with sample size reassessment for promising interim results

12:30 – 13:30 **LUNCH BREAK**

13:30 – 15:15 **SESSION 2: MULTIPLE TESTING (1)**
CHAIR: WILLI MAURER (BASEL), MARTIN POSCH (LONDON)

Ajit Tamhane, Dror Rom (Evanston/USA): An improved Hommel-Hochberg hybrid procedure

Gerhard Hommel (Mainz): p-values are random variables – are they really?

Klaus Strassburger, Helmut Finner (Düsseldorf): Randomized p-values and randomized empirical distribution functions in multiple testing

Kornelius Rohmeyer (Hannover), Florian Klinglmüller (Wien): gMCP – an R package for graphical multiple test problems

Kathrin Stucke, Meinhard Kieser (Heidelberg): Sample size calculation for three-arm non-inferiority trials with Poisson distributed count data

SCIENTIFIC PROGRAM – DETAILED TIME SCHEDULE

15:15 – 15:45 **COFFEE BREAK**

15:45 – 16:45 **INVITED SESSION**
CHAIR: MEINHARD KIESER (HEIDELBERG)

Sue-Jane Wang (Silver Spring/USA): Adaptive designed clinical trials and their associated multiplicity issues including FDA's currently thinking and perspectives

Hsien-Ming James Hung (Silver Spring/USA): Statistical considerations and multiplicity issues in active control trial designs

16:45 – 17:00 **COFFEE BREAK**

17:00 – 18:45 **SESSION 3: SEAMLESS PHASE II/III DESIGNS AND DOSE-FINDING**
CHAIR: STEFAN ENGLERT (HEIDELBERG), JAMES HUNG (SILVER SPRING/USA)

Cornelia Ursula Kunz (Warwick), Tim Friede (Göttingen): Adaptive treatment selection in seamless phase II/III trials using short-term endpoints

Lisa Hampson (Lancaster), Christopher Jennison (Bath): Optimal data combination rules in seamless phase II/III clinical trials

Maximo Carreras (Basel), Georg Gutjahr (Bremen): Seamless phase II/III adaptive designs with treatment selection based on drug exposure, toxicity and response

Alexandra Graf, Peter Bauer (Wien): Maximum type I error rate inflation in multi-armed clinical trials with interim sample size modifications

Georg Gutjahr (Bremen), Björn Bornkamp (Basel): MCP-mod without guesstimates

19:15 – 20:00 **GUIDED TOUR**

20:00 **CONFERENCE DINNER**

SCIENTIFIC PROGRAM – DETAILED TIME SCHEDULE

FRIDAY, JULY 6

08:30 – 10:15 **SESSION 4: TIME-TO-EVENT DATA AND CONFIDENCE INTERVALS**
CHAIR: WERNER BRANNATH (BREMEN), GERNOT WASSMER (KÖLN)

Sandra Ligges (Münster), Gernot Wassmer (Köln): Estimation of the hazard ratio in adaptive designs with sample size readjustment

Sebastian Irle, Helmut Schäfer (Marburg): Interim design modifications in time-to-event studies

Rene Schmidt, Joachim Gerss (Münster): Two-stage adaptive designs with test statistics with arbitrary dependence structure based on the inverse normal method

Dominic Magirr, Thomas Jaki (Lancaster): Simultaneous confidence intervals that are compatible with closed testing in adaptive designs

Sylvia Schmidt, Werner Brannath (Bremen): Informative simultaneous confidence intervals

10:15 – 10:30 **COFFEE BREAK**

10:30 – 12:15 **SESSION 5: MULTIPLE TESTING (2)**
CHAIR: GERHARD HOMMEL (MAINZ), AJIT TAMHANE (EVANSTON/USA)

Jens Stange, Thorsten Dickhaus (Berlin): An effective number of tests

Marsel Scheer (Düsseldorf): Exceedance control of the number of false rejections in multiple testing

Thorsten Dickhaus, Jakob Gierl (Berlin): Simultaneous test procedures in terms of p-value copulae

Eric Derobert, Julie Perez (Paris): A parametrized strategy of gatekeeping, keeping untouched the probability of having at least one significant result

Geraldine Rauch, Meinhard Kieser (Heidelberg): Multiplicity adjustment for composite binary endpoints

12:15 – 12:30 **COFFEE BREAK**

Scientific Program – Detailed Time Schedule

12:30 – 14:15 **Session 6: Selection of Treatments or Populations**
Chair: Tim Friede (Göttingen), Helmut Schäfer (Marburg)

Gernot Wassmer, Silke Jürgens (Köln): Designing issues in population enrichment designs

Ekkehard Glimm (Basel): Clinical trial designs with delayed selection of the primary comparison

James Wason, Jack Bowden (Cambridge): Multi-stage drop-the-loser designs

Jack Bowden (Cambridge), Ekkehard Glimm (Basel): Conditionally unbiased and near unbiased estimation for multi-stage drop-the-losers designs

Matthew Sydes, Mahesh Parmar (London): Flexible trial design in practice. Stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: a multi-arm multi-stage randomised controlled trial

14:15 **Meeting of Working Group "Adaptive Designs and Multiple Testing Procedures" of IBS-DR and ROeS**

Abstracts

SESSION 1 (1), JULY 5, 10:30 – 12:30

BLINDED AND UNBLINDED INTERNAL PILOT STUDY DESIGNS FOR CLINICAL TRIALS WITH OVERDISPERSED COUNT

Simon Schneider
University Medical Center Göttingen
simon.schneider@med.uni-goettingen.de

Heinz Schmidli
Novartis Pharma AG

Tim Friede
University Medical Center Göttingen

In the planning phase of a clinical trial with counts as primary outcomes, such as relapses in Multiple Sclerosis (MS), there is uncertainty with regard to the nuisance parameters (e.g. overall event rate, the dispersion parameter) which need to be specified for sample size estimation. For this reason the application of adaptive designs with blinded sample size reestimation (BSSR) are attractive (Cook et al. 2009, Friede and Schmidli 2010a). After a comparison of existing methods we consider in this presentation a modified version of the maximum likelihood method for BSSR for negative binomial data proposed by Friede and Schmidli (2010b). The method works well in terms of sample size distribution and power, if the assumed clinically effect is equal to the true effect. We compare the BSSR approach to an unblinded procedure in situations where an uncertainty about the assumed effect size exists. For practically relevant scenarios we make recommendations when application of the blinded or unblinded procedure are indicated. In addition, results for unbalanced designs previously not considered are shown in a simulation study. The methods are illustrated by a study in Relapsing Remitting MS.

References:

[1] Cook, RJ. et al.. (2009). Two-stage design of clinical trials involving recurrent events. Statistics in Medicine, 28: 2617-2638.
[2] Friede, T. and Schmidli, H. (2010). Blinded sample size reestimation with count data: Methods and applications in multiple sclerosis. Statistics in Medicine, 29: 1145-1156.
[3] Friede, T. and 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.

SESSION 1 (2), JULY 5, 10:30 – 12:30

ADAPTIVE SAMPLE SIZE RE-ESTIMATION FOR RECURRENT EVENT DATA

Katharina Ingel
University Medical Center Mainz
katharina.ingel@unimedizin-mainz.de

Antje Jahn-Eimermacher
University Medical Center Mainz

Some clinical trials compare the repeated occurrence of the same type of event, e.g. epileptic seizures or acute otitis media, between two or more treatment groups. The Andersen-Gill model has been proposed to analyse such recurrent event data and applies a robust variance estimate to control the type I error [1].

For sample size calculation, Bernardo and Harrington suggest a formula, which relies on homogeneity in patients' baseline hazard [2]. If this assumption is violated, the actual power of the trial will be decreased.

We adjust the sample size formula to achieve the anticipated power even if the patients are heterogeneous in their baseline hazard. For this purpose, we introduce a nuisance parameter, which is derived from characteristics of the robust variance estimate [3] and depends on the degree of heterogeneity, the duration of follow-up, the baseline-hazard and the treatment-effect. Some of these parameters will usually be unknown in the planning phase of a trial. We explore how an adaptive sample size adjustment can be used to estimate the nuisance parameter and subsequently re-estimate the sample size. The interim analysis will be performed after the number of events, which is required when assuming homogeneous baseline hazards, is observed. The more heterogeneity the interim analysis reveals the higher the final sample size will be.

The performance of this internal sample size re-estimation design with respect to type I error and power is evaluated by the use of simulations. We illustrate our results with clinical data.

References:

[1] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. The Annals of Statistics, 10:1100-1120.
[2] Bernardo, M.V.P. and Harrington, D.P. (2001). Sample size calculations for the two-sample problem using the multiplicative intensity model. Statistics in Medicine, 20:557-579.
[3] Al-Khalidi, H.R., Hong, Y., Fleming, T.R. and Therneau, T.M. (2011). Insights on the robust variance estimator under recurrent-events model. Biometrics, 67:1564-1572.

SESSION 1 (3), JULY 5, 10:30 – 12:30

BLINDED CONTINUOUS MONITORING OF THE NUISANCE PARAMETER IN CLINICAL TRIALS

Frank Miller
AstraZeneca
frank.miller@astrazeneca.com

Tim Friede
University Medical Center Göttingen

Determination of a clinical trial's size is an important task in the planning of any trial because of the direct implications of the sample size on feasibility, costs and timelines. However, sample size calculations are often subject to substantial uncertainty due to limited prior information on the size of nuisance parameters such as variances or event rates. Continuous monitoring of the nuisance parameter in clinical trials has been proposed as a tool to size trials appropriately. With this approach, the nuisance parameter is continuously monitored during the trial. The trial is stopped when the actual estimate for the nuisance parameter and sample size fulfil a stopping criterion. Continuous monitoring can therefore be viewed as stochastic process with stopping time.

In this presentation, we describe the bias that occurs in unblinded continuous monitoring of the variance by means of a simulation study. Then we propose a procedure for blinded continuous monitoring that would not require breaking the treatment code during the on-going study and show that the procedure does not suffer from the same biases observed in unblinded monitoring. Results on the performance properties of such designs are given and the designs are compared to blinded reestimation procedures with a single data look. Furthermore, we present a hypertension trial where blinded sample size reestimation with a single data look was applied and investigate the properties of blinded continuous monitoring in this setting. Finally, we close with a brief discussion.

SESSION 1 (4), JULY 5, 10:30 – 12:30

EVALUATION OF SAMPLE SIZE ADAPTATION RULES IN CLINICAL STUDIES AIMING AT AN OVERALL PERFORMANCE OPTIMIZATION

Stefan Englert
University of Heidelberg
englert@imbi.uni-heidelberg.de

Meinhard Kieser
University of Heidelberg

Adaptive designs have been widely used in clinical practice. Such designs allow for mid-course design modifications, for example to adjust the sample size based on the observed treatment effect at interim. Usually, these designs are initially planned as fixed group-sequential designs, i.e. with sample sizes at each stage fixed in advance. In a next step, design characteristics of specific recalculation rules are investigated and an optimal strategy, for example, with respect to average or total sample size is selected.

In this talk, we present which designs result if both steps are not considered in isolation, but if the initial design is optimized for the planned adaptation rule. In this case, the overall optimization accounts for both the initial design and the applied recalculation strategy. Especially, designs with discrete test statistics are considered. In these designs, an exhaustive search over all possible sample sizes choices conditional on the interim results is possible. Therefore, conditional on the interim result the sample size can be chosen which is optimal with respect to the selected optimality criteria. In the calculations, the branch-and-bound algorithm was used to keep the computation effort feasible.

The impact of these 'optimal' rules for adjusting the sample sizes in clinical studies using flexible two-stage designs is shown for a variety of constellations. It is demonstrated that a combined optimization of both the planning and reassessment strategy may lead to counterintuitive recalculation rules and design features.

References:

[1] Bauer, P. and König, F. (2006). The reassessment of trial perspectives from interim data — a critical view. Statistics in Medicine, 25:23-36.

SESSION 1 (5), JULY 5, 10:30 – 12:30

TESTING PRIMARY AND SECONDARY ENDPOINTS IN ADAPTIVE DESIGNS WITH SAMPLE SIZE REASSESSMENT FOR PROMISING INTERIM RESULTS

Florian Klinglmüller
Medical University Vienna
florian.klinglmueller@meduniwien.ac.at

Franz König
Medical University Vienna

Lingyung Liu
Cytel, Harvard School of Public Health

Cyrus Mehta
Cytel, Harvard School of Public Health

Recently adaptive designs attracted much interest where the sample size is increased for promising interim results of the primary endpoint still using the conventional test statistic [3]. Complex multiple testing strategies for testing primary and secondary endpoints have been extensively discussed for fixed sample designs, only few publications deal with this issue in the group sequential setup [2, 4].

We investigate the impact on the multiple type I error rate when hierarchically testing primary and secondary endpoints using adaptive designs with sample size reassessment using conditional power arguments for the primary endpoint. Different testing strategies will be evaluated. E.g., if the primary endpoint is rejected, the secondary endpoint will be tested using the conventional pooled test statistic (ignoring the adaptive nature of the trial). This will be compared to other adaptive methods. Extending the work of [3] and [1] we define promising zones for both primary and secondary endpoints, where the type I error rate will be strictly controlled if a certain type of sample size increase is performed.

References:

[1] Brannath, W. and Koenig, F. (2007). Multiplicity and flexibility in clinical trials. Pharmaceutical Statistics, 6:205-216.
[2] Glimm, E., Maurer, W. and Bretz, F. (2010). Hierarchical testing of multiple endpoints in group-sequential trials. Statistics in Medicine, 29:219-228.
[3] Mehta, C. R. and Pocock, S. J. (2010). Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statistics in Medicine, 30:3267-3284.
[4] Tamhane, A. C., Mehta, C. R. and Liu, L. (2010). Testing a primary and a secondary endpoint in a group sequential design. Biometrics, 66:1174-1184.

SESSION 2 (1), JULY 5, 13:30 – 15:15

AN IMPROVED HOMMEL-HOCHBERG HYBRID PROCEDURE

Ajit Tamhane
Northwestern University
atamhane@northwestern.edu

Dror Rom
Prosoft Software, Inc.

Jiangtao Gou
Northwestern University

Dong Xi
Northwestern University

Let $p_1, p_2, \ldots, p_n$ be independent p-values associated with null hypotheses $H_1, H_2, \ldots, H_n$ and let $p_{(1)}, p_{(2)}, \ldots, p_{(n)}$ be the ordered p-values with $H_{(1)}, H_{(2)}, \ldots, H_{(n)}$ the corresponding null hypotheses. We want to test $H_1, H_2, \ldots, H_n$ at the familywise error rate (FWER) level α. Hochberg's (1988) procedure uses a simple step-up algorithm which tests the hypotheses beginning with $H_{(n)}$ and stops and rejects all remaining hypotheses in the sequence if at step i, $p_{(n-i+1)} < \alpha/i$. Hommel's (1988) procedure is more powerful than Hochberg's but uses a more complicated algorithm: if at step i, $p_{(n-j+1)} < (i-j+1)\alpha/i$ for at least one $j = 1, \ldots, i$ then it rejects all hypotheses with $p_j < \alpha/(i-1)$. We propose a procedure that combines the simple step-up algorithm of Hochberg with Hommel's rejection rule: if at step i, $p_{(n-i+1)} < c_i\alpha$ and then reject all hypotheses with $p_j < \alpha/i$. We show that this procedure controls the FWER if $c_i = (i+1)\alpha/2i$; more exact values can be numerically computed of which these values are limits. Power comparisons are made via simulation with competing procedures including that of Rom (1990).

References:

[1] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple significance testing, Biometrika. 75: 800-802.
[2] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika, 75:383-386
[3] Marcus, R., Peritz, E. and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika, 63: 655-660.
[4] Rom, D. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika, 77:663-665.
[5] Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. Biometrika, 73: 751-754.

SESSION 2 (2), JULY 5, 13:30 – 15:15

P-VALUES ARE RANDOM VARIABLES - ARE THEY REALLY?

Gerhard Hommel
University Medical Center Mainz
gerhard.hommel@unimedizin-mainz.de

A common definition for a p-value is that it is the smallest significance level at which the null hypothesis can be rejected. Another property of a p-value is that, under the null hypothesis, its distribution is stochastically larger than or equal to the uniform distribution on [0;1]. This property is utilized, in particular, for the construction of many multiple test procedures or for adaptive designs (p-clud condition). However, this means that the p-value can be considered as a random variable.

In my talk, I will show that this assumption is justified when the null hypothesis consists of only a countably infinite number of parameter points or when the p-value can be generated by at most a countably infinite number of tests. On the other hand, when this number is uncountably infinite, one can construct a p-value that is no more a random variable. I present an example of such a construction which can, however, never occur in practice.

SESSION 2 (3), JULY 5, 13:30 – 15:15

RANDOMIZED P-VALUES AND RANDOMIZED EMPIRICAL DISTRIBUTION FUNCTIONS IN MULTIPLE TESTING

Klaus Strassburger
German Diabetes Center at Heinrich-Heine-University Düsseldorf
Leibniz Institute for Diabetes Research
strass@ddz.uni-duesseldorf.de

Helmut Finner
German Diabetes Center at Heinrich-Heine-University Düsseldorf
Leibniz Institute for Diabetes Research

Discrete test statistics may cause additional issues in multiple hypotheses testing problems. For example, classical p-values derived from discrete test statistics are typically stochastically larger than a uniform variate if the corresponding hypotheses are true. It is known (see for example Finner et al. 2010) that this draw back can lead to very conservative multiple test procedures, especially when these procedures are based on the empirical distribution function (ecdf) of such discrete p-values. One way out of this dilemma are realized randomized p-values, cf. Finner and Strassburger (2007). However, the use of realized randomized p-values can be criticized because the final decision is not fully determined by the observed data since it depends on extra randomization experiments. To overcome this concern we propose to construct non-randomized multiple tests based on the (formally) randomized ecdf of the p-values which is defined as the conditional expectation of the ecdf of the realized randomized p-values with respect to the extra randomization experiments. This approach seems to work well, but it is yet unclear whether properties of the realized randomized version of the test procedure, such as the control of a given error criteria (e.g. FWER or FDR) carry over to its non-randomized counterpart. We will shed some light on this question, by investigating plug-in procedures involving ecdf-based estimates of the proportion of true hypotheses.

References:

[1] Finner, H. and Strassburger, K. (2007). A note on p-values for two-sided tests. Biometrical Journal, 49:941-943.
[2] Finner, H., Strassburger, K., Heid, I.M., Herder, C., Rathmann, W., Giani, G., Dickhaus, T., Lichtner, P., Meitinger, T., Wichmann, HE., Illig, T., Gieger, C. (2010). How to link call rate and p-values for Hardy-Weinberg equilibrium as measures of genome-wide SNP data quality. Statistics in Medicine, 29:2347-2358.

SESSION 2 (4), JULY 5, 13:30 – 15:15

GMCP - AN R PACKAGE FOR GRAPHICAL MULTIPLE TEST PROCEDURES

Kornelius Rohmeyer
Leibniz University of Hannover
rohmeyer@small-projects.de

Florian Klinglmueller
Medical University of Vienna

In multiple testing problems the relations and priorities between elementary hypotheses often can be adequately described by a weighted graph as Bretz et al. (2009) have shown. Each graph defines a weighting strategy for the set of elementary hypotheses and each of its subsets. This leads for example to corresponding weighted Bonferroni-based, Simes-based or parametric closed test procedures that control the familywise error rate.

With the open source R package gMCP we provide a framework and Java based graphical user interface to design appropriate weighted graphs for test problems or to choose alternatively from many examples from the literature. In addition to the mentioned tests also adjusted p-values and in some cases even compatible simultaneous confidence intervals can be calculated. Power analysis tools are provided and the import/export of graphs, data and settings helps with the creation of reports. With all functionality available from a graphical user interface as well as from the R command line, the package is designed both for people without any background in R as well as for R experts.

The talk will give an overview of the package and explain helpful features with examples from the literature.

References:

[1] Bretz, F., Maurer, W., Brannath, W. and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine, 28:586-604
[2] Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W. and Rohmeyer K. (2011). Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes or parametric tests. Biometrical Journal, 53:894-913
[3] Rohmeyer, K., Klinglmueller, F. and Bornkamp, B. (2012). gMCP: Graph based multiple comparison procedures, R package version 0.7-8.
URL: http://CRAN.R-project.org/package=gMCP/

SESSION 2 (5), JULY 5, 13:30 – 15:15

SAMPLE SIZE CALCULATION FOR THREE-ARM NON-INFERIORITY TRIALS WITH POISSON DISTRIBUTED COUNT DATA

Kathrin Stucke
University of Heidelberg
stucke@imbi.uni-heidelberg.de

Meinhard Kieser
University of Heidelberg

In the three-arm 'gold standard' non-inferiority design an experimental treatment, an active reference and a placebo are compared. This design is becoming increasingly popular and is, whenever feasible, recommended for use by regulatory guidelines. However, comparatively few research has been done on the topic of sample size calculation for studies with this design yet. Clinical trials with count data as the primary outcome are common in various medical areas. Examples are relapse counts in multiple sclerosis trials or the number of attacks in trials for the treatment of migraine. We present a method for sample size calculation and identification of an optimal sample size allocation for the three-arm 'gold standard' design with Poisson distributed count data. It turns out that optimal allocation may lead to a considerable decrease of the total sample size and assigns in many situations more patients to the active treatment groups than to placebo.

We apply our method to a recently published clinical trial in multiple sclerosis and show how the optimal allocation ratio can be used as a starting point to determine a sample size assignment rule that is favourable both from statistical and practical viewpoints.

INVITED SESSION, JULY 5, 15:45 – 16:45

ADAPTIVE DESIGNED CLINICAL TRIALS AND THEIR ASSOCIATED MULTIPLICITY ISSUES INCLUDING FDA'S CURRENTLY THINKING AND PERSPECTIVES

Sue-Jane Wang
Food and Drug Administration, Silver Spring, MD, U.S.A.
Suejane.Wang@fda.hhs.gov

STATISTICAL CONSIDERATIONS AND MULTIPLICITY ISSUES IN ACTIVE CONTROL TRIAL DESIGNS

Hsien-Ming James Hung
Food and Drug Administration, Silver Spring, MD, U.S.A.
HsienMing.Hung@fda.hhs.gov

SESSION 3 (1), JULY 5, 17:00 – 18:45

ADAPTIVE TREATMENT SELECTION IN SEAMLESS PHASE II/III TRIALS USING SHORT-TERM ENDPOINTS

Cornelia Ursula Kunz
Warwick Medical School
c.u.kunz@warwick.ac.uk

Tim Friede
University Medical Center Göttingen

Nicholas Parsons
Warwick Medical School

Susan Todd
Department of Mathematics and Statistics, University of Reading

Nigel Stallard
Warwick Medical School

Adaptive seamless phase II/III designs with treatment selection at an interim analysis have become increasingly more attractive due to their potential to save development costs and to shorten time-to-market of a new treatment. Different methods have been proposed for selection of the treatment group that will continue along with the control group to the second stage of the trial. If the primary endpoint is observed only after long-term follow-up it may be desirable to use short-term endpoint data at the interim analysis to select a treatment [1]. In other cases, at least some long-term endpoint data might be available at the time of the interim analysis which might be used together with the short-term endpoint data to estimate the treatment effect upon which the treatment selection can be based [2]. While appropriate methods to combine data from different stages of the trial ensure control of the family-wise type I error rate in the strong sense, the power of the different approaches for treatment selection differs depending on several assumptions. In this talk we present the results of a formal comparison of different methods for treatment selection based on analytical results and a simulation study, together with a summary of the strengths and weaknesses of the approaches. Based on these results, we show how existing methods can be improved, increasing both the probability of selecting the most effective treatment and the power.

References:

[1] Friede, T., Parsons, N., Stallard, N., Todd, S., Valdes Marquez, E., Chataway, J. and Nicholas, R. (2011). Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. Statistics in Medicine, 30:1528-1540.
[2] Stallard, N. (2010). A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. Statistics in Medicine, 29:959-971.

SESSION 3 (2), JULY 5, 17:00 – 18:45

OPTIMAL DATA COMBINATION RULES IN SEAMLESS PHASE II/III CLINICAL TRIALS

Lisa V. Hampson
Lancaster University
l.v.hampson@lancaster.ac.uk

Christopher Jennison
University of Bath

We consider seamless Phase II/III clinical trials, which compare K treatments against a common control in stage 1 and select the most promising for further testing against control in stage 2. Such a trial requires careful upfront planning if it is to win regulatory acceptance as a pivotal study. For seamless trials to be attractive, this increased planning should be offset by efficiency gains made possible because data accumulated across the study are combined to make a final decision on the efficacy of the selected treatment. We derive optimal versions of final decision rules maximising power. This is a multivariate decision problem because properties of rules depend on a vector of means.

Rules with the correct familywise error rate maximising power for different configurations of means are found as solutions to Bayes decision problems. Different solutions are found as the shape of the mean vector changes but we find only small gains in power are possible by ma king strong assumptions about the structure of the mean vector. By studying procedures with optimal decision rules, we assess the efficiency of alternative proposals, namely closed testing procedures based on p-value combination rules, and rules using only data on the selected treatment and control for final decisions. For procedures with efficient decision rules, we find that Phase II observations on the selected treatment and control retain between 22-98% of their value as Phase III observations. Thus, efficient seamless designs can offer large savings in sample size which may have important implications, for example, for the feasibility of trials in rare diseases.

SESSION 3 (3), JULY 5, 17:00 – 18:45

SEAMLESS PHASE II/III ADAPTIVE DESIGNS WITH TREATMENT SELECTION BASED ON DRUG EXPOSURE, TOXICITY AND RESPONSE

Maximo Carreras
F. Hoffmann-La Roche AG Basel
maximo.carreras@roche.com

Georg Gutjahr
University of Bremen

Werner Brannath
University of Bremen

The planning of an oncology clinical trial with a seamless phase II/III adaptive design is discussed. Two regimens of an experimental treatment are compared to a control at the end of phase II and the most-promising regimen is selected to continue, together with control, into phase III. Since the study's primary endpoint, overall survival (OS), will be immature at the time of regimen-selection analysis, it is of interest to investigate whether the incorporation of surrogate information such as drug exposure and toxicity can help improving the regimen-selection process and thus the study's probability of success. To this end, designs are considered which include the primary as well as surrogate endpoints in the regimen-selection analysis. At the end of the study, testing of efficacy is carried out to compare the selected regimen to the control with respect to OS, utilizing relevant data from both phases.

Several approaches for testing the primary hypothesis are assessed with regards to power and type I error rate. Since the operating characteristics of these designs depend on the specific regimen-selection rules considered, benchmark scenarios are proposed in which a perfect surrogate and no surrogate is used at the regimen-selection analysis. The operating characteristics of these benchmark scenarios provide a range where those of the actual study design are expected to lie.

A discussion on family-wise error rate control for testing primary and key secondary endpoints as well as an assessment of bias in the final treatment effect estimate for the selected regimen are also presented.

SESSION 3 (4), JULY 5, 17:00 – 18:45

MAXIMUM TYPE 1 ERROR RATE INFLATION IN MULTI-ARMED CLINICAL TRIALS WITH INTERIM SAMPLE SIZE MODIFICATIONS

Alexandra Graf
Medical University of Vienna
alexandra.graf@meduniwien.ac.at

Peter Bauer
Medical University of Vienna

Franz Koenig
Medical University of Vienna

Sample size modifications in an adaptive interim analysis based on the observed interim effects can considerably inflate the type 1 error rate if the pre-planned conventional fixed sample-size tests are applied in the final analysis, ignoring the adaptive character of the study. For a single treatment-control comparison Graf and Bauer (2011) have shown that if the allocation rate to treatment arm is modified after an interim analysis, the maximum inflation of the type 1 error rate may be substantially larger than in the case of sample size reassessment with stage-wise balanced sample sizes derived by Proschan and Hunsberger (1998). We investigate scenarios where more than one treatment arms are compared to a single control as well as scenarios with interim treatment selection by carrying on only the treatment with the largest observed interim effect and the control to the second stage. It is assumed that either a naive testing procedure with a conventional fixed sample-size test or a multiplicity adjusted Dunnett test is performed in the final analysis. The maximum inflation of the type 1 error rate for such types of design can be calculated by searching for "worst case" scenarios, i.e. sample size adaptation rules that lead to the largest conditional type 1 error rate in any point of the sample space. To achieve the maximum type 1 error rate, we first assume unconstrained second-stage-sample-sizes. To see how the numbers will change in more realistic scenarios, we put constraints on the second-stage-sample-size, which may lead to scenarios not inflating the type 1 error rate.

References:

[1] Graf, AC., and Bauer, P. (2011). Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look. Statistics in Medicine, 30: 1637-1647.
[2] Proschan, M.A. and Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. Biometrics, 51: 1315-1324.

SESSION 3 (5), JULY 5, 17:00 – 18:45

MCP-MOD WITHOUT GUESSTIMATES

Georg Gutjahr
University of Bremen
georg.gutjahr@math.uni-bremen.de

Bjorn Bornkamp
Novartis Pharma AG

In the MCP-mod approach (Bretz et al., 2005), dose-finding studies are analysed by a combination of modelling and multiple comparison procedures. Before the start of the trial, on selects a number of possible dose-response shapes, such as the Emax or the logistic model. Then, after the trial, the parameters of the possible dose-response shapes are estimated in the modelling part. In the testing part, linear contrast tests are used to reject the hypothesis that the dosis has no influence on the responses (proof-of-concept). The weights in the contrast tests are pre-specified based on guesstimates about the parameter values in each of the possible dose-response shapes. In practice, elicitation of such guesstimates before the start of the trial is difficult; if the guesstimates are not sufficiently accurate, the power of the resulting test procedure will decrease. Therefore, it would be desirable to use the actual parameter estimates from the modelling part in place of the guesstimates. This talk will describe a modification of the linear contrast tests to account for the resulting data-dependent weights.

References:

[1] Bretz F., Pinheiro J., Branson M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. Biometrics, 61, 738-748.

SESSION 4 (1), JULY 6, 08:30 – 10:15

ESTIMATION OF THE HAZARD RATIO IN ADAPTIVE DESIGNS WITH SAMPLE SIZE READJUSTMENT

Sandra Ligges
Universitätsklinikum Münster
sandra.ligges@ukmuenster.de

Gernot Wassmer
Aptiv Solutions Cologne

Christine Müller
Technische Universität Dortmund

In adaptive designs stagewise independent data is crucial for the validity of the performed procedures. Independency is usually achieved by dealing with different patient collectives. This is not possible in survival studies where patients may contribute information to subsequent stages. Here the conventional test statistics have to be modified in order to retain independent stagewise inference. In the literature two different strategies have been pursued. Independent information can be obtained by either using increments of certain pivot statistics, e.g. logrank statistics [3], or by right-censoring and left-truncating the data at the time points of the interim analyses [2].

It was aimed to construct estimators for the hazard ratio in two-armed two-stage adaptive designs with survival endpoints and sample size readjustment at the interim analysis. This was achieved by firstly utilising either of these two introduced methods for splitting up information into two independent parts and by secondly following the general construction principle proposed by [1]. A simulation study for the comparison of the different generated estimators in various scenarios of adaptive designs relying on inverse normal type boundaries for increments in logrank statistics [3] was carried out. In the present talk these estimators and some results of the simulation study will be presented.

References:

[1] Brannath, W., König, F., and Bauer, P. (2006). Estimation in flexible two stage designs. Statistics in Medicine, 25:3366-3381.
[2] Jahn-Eimermacher, A. and Ingel, K. (2009). Adaptive trial design: A general methodology for censored time to event data. Contemporary Clinical Trials, 30:171-177.
[3] Wassmer, G. (2006). Planning and analyzing adaptive group sequential survival trials. Biometrical Journal, 48:714-729.

SESSION 4 (2), JULY 6, 08:30 – 10:15

INTERIM DESIGN MODIFICATIONS IN TIME-TO-EVENT STUDIES

Sebastian Irle
IMBE Marburg
irle@staff.uni-marburg.de

Helmut Schäfer
IMBE Marburg

We propose a flexible method for interim design modifications in time-to-event studies. With this method, it is possible to inspect the data at any time during the course of the study, without the need for pre-specification of a learning phase, and to make certain types of design modifications depending on the interim data without compromising the type I error risk. The method can be applied to studies designed with a conventional statistical test, fixed sample or group sequential, even when no adaptive interim analysis and no specific method for design adaptations (such as combination tests) had been foreseen in the protocol. Currently, the method supports design changes such as an extension of the recruitment or follow-up period, as well as certain modifications of the number and the schedule of interim analyses as well as changes of inclusion criteria. In contrast to existing methods offering the same flexibility, our approach allows to make use of the full interim information collected until the time of the adaptive data inspection. This includes time-to-event data from patients who have already experienced an event at the time of the data inspection, and preliminary information from patients still alive, even if this information is predictive for survival, such as early treatment response in a cancer clinical trial.

Our method is an extension of the so-called conditional rejection probability (CRP) principle. It is based on the conditional distribution of the test statistic given the final value of the same test statistic from a subsample, namely the learning sample. It is developed in detail for the example of the log-rank statistic, for which we derive this conditional distribution using martingale techniques.

References:

[1] Irle, S. and Schäfer, H. (in press) Interim design modifications in time-to-event studies. Journal of the American Statistical Association.

SESSION 4 (3), JULY 6, 08:30 – 10:15

TWO-STAGE ADAPTIVE DESIGNS WITH TEST STATISTICS WITH ARBITRARY DEPENDENCE STRUCTURE BASED ON THE INVERSE NORMAL METHOD

Rene Schmidt
Westfälische Wilhelms-Universität Münster
rene.schmidt@ukmuenster.de

Joachim Gerss
Westfälische Wilhelms-Universität Münster

Andreas Faldum
Westfälische Wilhelms-Universität Münster

Adaptive designs were originally developed for independent test statistics. This is true for example if the data for each stage come from different units and are normally distributed with known variance. Another possibility to get independent test statistics is to exploit the independent increment structure of some statistical models. Sometimes it may not be possible to satisfy these conditions or to check whether they are satisfied. In these cases, the test statistics and p-values of each stage may be dependent. Depending on the design parameters and on the true dependence structure between the p-values of the stages, the decisions can become conservative as well as anticonservative. In general, there may be uncountable dependence structures. We investigate the type I error of two-stage adaptive designs if any dependence structure between the test statistics from the stages is assumed to be admissible (worst case scenario). For this purpose, we perform analytical considerations under the restriction that the conditional error function is given according to the inverse normal method. We discuss how the significance level of the unweighted inverse normal design is inflated in the worst case as compared to the situation of independent stages. On this basis the decision boundary for the second stage may be modified so that the type I error is controlled in the worst case and thus for any dependence structure.

SESSION 4 (4), JULY 6, 08:30 – 10:15

SIMULTANEOUS CONFIDENCE INTERVALS THAT ARE COMPATIBLE WITH CLOSED TESTING IN ADAPTIVE DESIGNS

Dominic Magirr
Lancaster University
d.magirr@lancaster.ac.uk

Thomas Jaki
Lancaster University

Martin Posch
European Medicines Agency

We describe a general method for finding a confidence region for a vector of K unknown parameters that is compatible with the decisions of a two stage closed testing procedure in an adaptive experiment. The closed test procedure is characterized by the fact that rejection or nonrejection of a null hypothesis may depend on the decisions for other hypotheses and the compatible confidence region will, in general, have a complex, nonrectangular shape. We find the smallest cross product of simultaneous confidence intervals containing the region and provide computational shortcuts for calculating the lower bounds for parameters corresponding to the rejected null hypotheses. An appealing property of these lower bounds is that they may provide more informative inference than the original closed test procedure despite failure to reject all individual null hypotheses. This is in contrast to related methods for fixed sample experiments. We illustrate the methodology with the example of an adaptive Phase II/III clinical trial.

SESSION 4 (5), JULY 6, 08:30 – 10:15

INFORMATIVE SIMULTANEOUS CONFIDENCE INTERVALS

Sylvia Schmidt
University of Bremen
sylviaschmidt@math.uni-bremen.de

Werner Brannath
University of Bremen

In multiple testing, one often wishes not only to have great power for rejecting hypotheses but also to obtain additional information through simultaneous confidence intervals (SCIs). While single-step tests like Bonferroni offer a canonical way to construct SCIs, this is not obvious for the more powerful stepwise tests like Bonferroni-Holm. The methods proposed in Strassburger and Bretz (2008) and Guilbaud (2008, 2009) lead to consistent SCIs for a broad class of multiple tests. However, in some cases these SCIs are not informative for rejected hypotheses, i.e., they contain all parameters of the alternative.

We consider multiple test procedures for one-sided hypotheses $H_i: \theta_i \leq 0$, $i = 1, \dots, m$, and SCIs (L_i, ∞), where L_i are bounds defined in terms of individual p-values for the m hypotheses. The Bonferroni-Holm procedure and its extension to SCIs improve the Bonferroni test and SCIs uniformly in the sense that $L_i^{Bonf} \geq 0$ implies $L_i^{Holm} \geq 0$, but it is inferiour with respect to informative rejection, i.e., $L_i^{Holm} > 0$ implies $L_i^{Bonf} > 0$. We propose a method to construct SCIs that uniformly improve the Bonferroni SCIs with respect to informative rejection. Our test does not reject in all cases where Holm does, but it has higher power than Bonferroni and produces always informative SCIs for all hypotheses.

The new approach is based on the projection method, where p-values $p(\delta)$ for all δ in the parameter space are defined so that the confidence domain consists of all parameters with p-value larger thanα. Then the projection of the confidence domain results in SCIs that have a coverage probability of at least $1 - \alpha$. Our p-values $p(\delta)$ will be adjusted p-values of weighted Bonferroni tests with parameter-dependent weights involving continuous penalizing functions $\lambda_i(\delta_i)$. The choice of the λ_i gives the flexibility to put more emphasis on smaller values of δ_i which makes them easier to reject and therefore to obtain informative SCIs.

We will present our method and compare it to existing approaches. Thanks to a numerical result, we obtain an easy algorithm to implement this method. We show results from a simulation study comparing our approach to the Bonferroni and the Bonferroni-Holm procedure with respect to power and informative rejections. Extensions to other classes of test procedures like general union-intersection tests and hierarchical tests will be outlined.

References:

[1] Guilbaud, O. (2008). Simultaneous confidence regions corresponding to Holm's step-down procedure and other closed-testing procedures. Biometrical Journal, 50:678-692.
[2] Guilbaud, O. (2009). Alternative confidence regions for Bonferroni-based closed-testing procedures that are not alpha-exhaustive. Biometrical Journal, 51:721-735.
[3] Strassburger, K. and Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. Statistics in Medicine, 27:4914-4927.

SESSION 5 (1), JULY 6, 10:30 – 12:15

AN EFFECTIVE NUMBER OF TESTS

Jens Stange
Humboldt-Universität zu Berlin
stange@math.hu-berlin.de

Thorsten Dickhaus
Humboldt-Universität zu Berlin

We consider a special class of multiple testing problems, consisting of M simultaneous point hypothesis tests in local statistical experiments. In other words, we restrict attention to two-sided alternatives in each marginal. Under certain structural assumptions the intersection overall M hypotheses (i. e., the global hypothesis) contains exactly one element ϑ^* (say), and it is easy to verify that the subset pivotality condition (see [1]) holds true. Moreover, ϑ^* is least favourable parameter configuration with respect to the familywise error rate (FWER) of single-step tests, meaning that the FWER of such tests becomes largest under ϑ^*.

Furthermore, it turns out that concepts of positive dependence are applicable to the involved test statistics in many practically relevant cases, e. g., for multivariate normal distributions, multivariate t-, F-distributions, or certa in classes of multivariate Γ-distributions (cf. [2,3]). This allows for a relaxation of the adjustment for multiplicity by making use of the intrinsic correlation structure in the data. In particular, conditional strong positive orthant dependence (CSPOD) (see [4]) leads to the computation of an "effective number of tests". Combining all this, we deduce a bound for the FWER in terms of a relaxed Sidak correction of the overall significance level.

Our findings can be applied to a variety of simultaneous location parameter problems, as in ANOVA-models or in the context of simultaneous categorical data analysis. For example, simultaneous chi-square tests for association of categorical features are ubiquitous in genomewide association studies with case-control setup (association between many single nucleotide polymorphisms and a binary phenotype). In this type of model, Moskvina and Schmidt (see [5]) gave a formula for an effective number of tests utilizing Pearson's haplotypic correlation coefficient as a linkage disequilibrium measure. Their result follows as a corollary from our general theory.

References:

[1] Westfall, P.H. and Young, S.S. (1993). Resampling-based multiple testing: examples and methods for P-value adjustment. Wiley, New York.
[2] Sidak, Z. (1973). On probabilities of rectangles in multivariate student distributions: Their dependence on correlations. The Annals of Mathematical Statistics, 42:169-135.
[3] Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. Journal of Multivariate Analysis, 10:467-498.
[4] Holland, P.W. and Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models. The Annals of Statistics, 14:1523-1543.
[5] Moskvina, V. and Schmidt, K.M. (2008). On multiple-testing correction in genome-wide association studies. Genetic Epidemiology, 32:567-573.

SESSION 5 (2), JULY 6, 10:30 – 12:15

EXCEEDANCE CONTROL OF THE NUMBER OF FALSE REJECTIONS IN MULTIPLE TESTING

Marsel Scheer
Heinrich-Heine-Universität Düsseldorf
mscheer@ddz.uni-duesseldorf.de

Controlling the k-FWER at level $\alpha \in (0,1)$ means that the probability of rejecting at least k true null hypotheses is bounded by α, cf. [1,2,3]. Considering k as fixed may be viewed as unsatisfactory. In this talk, we allow k to depend on the unknown number n_1 of false null hypotheses. For example, it seems more appropriate to require $k = k(n_1)$ to be small (large) if the number of false null hypotheses is small (large). We present sufficient conditions such that the probability of rejecting at least $k(n_1)$ true null hypotheses is asymptotically less than or equal α.

References:

[1] Victor, N. (1982). Exploratory data analysis and clinical research. Methods of Information in Medicine, 21:53-54.
[2] Hommel, G. and Hoffmann, T. (1988). Controlled uncertainty. In: Bauer, P. et al. (Eds.): Multiple Hypothesenprüfung. Springer, Berlin, 154-161.
[3] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. Annals of Statistics, 33:1138-1154.

SESSION 5 (3), JULY 6, 10:30 – 12:15

SIMULTANEOUS TEST PROCEDURES IN TERMS OF P-VALUE COPULAE

Thorsten Dickhaus
Humboldt-University Berlin
dickhaus@math.hu-berlin.de

Jakob Gierl
Humboldt-University Berlin

At least since the work of K. R. Gabriel (1969, see [1]), a broad class of multiple comparison procedures, so-called simultaneous test procedures (STPs), is established in the statistical literature. Elements of an STP are a testing family $(\mathcal{H}, \mathcal{T})$, consisting of a set of null hypotheses and corresponding test statistics, and a common critical constant c_α. The threshold c_α with which each of the test statistics has to be compared is calculated under the (joint) intersection hypothesis of $\mathcal{H}$. Under certain structural assumptions on $(\mathcal{H}, \mathcal{T})$, the so-constructed STP provides strong control of the family-wise error rate at level α. More recently, Hothorn et al. (cf. [2] and references therein) developed a general method to construct STPs in the case of asymptotic (joint) normality of the family $\mathcal{T}$ of test statistics, and provided numerical solutions in R to compute c_α in such cases.

Here, we propose to look at the problem from a different perspective. We will show that c_α can equivalently be expressed by a quantile of the copula of the family of p-values (or, more precisely, of distributional transforms as defined in [4]) associated with $\mathcal{T}$, assuming that each of these p-values is marginally uniformly distributed on the unit interval under the corresponding null hypothesis. This will offer the opportunity to exploit the rich and growing literature on copula-based modelling of multivariate dependency structures for multiple testing problems and in particular for the construction of STPs in non-Gaussian situations. Specifically, we will explain how parametric families of copulae, as extensively studied in [3], can be used to model an unknown or only partially known dependency structure of the p-values.

References:

[1] Gabriel, K. R. (1969). Simultaneous test procedures - some theory of multiple comparisons. Annals of Mathematical Statistics, 40:224-250.
[2] Hothorn, T., Bretz, F. and Westfall, P. (2008). Simultaneous inference in general parametric models. Biometrical Journal, 50:346-363.
[3] Nelsen, R. B. (2006). An introduction to copulas. 2nd ed. Springer Series in Statistics. New York, NY: Springer.
[4] Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. Journal of Statistical Planning and Inference, 139:3921-3927.

SESSION 5 (4), JULY 6, 10:30 – 12:15

A PARAMETRIZED STRATEGY OF GATEKEEPING, KEEPING UNTOUCHED THE PROBABILITY OF HAVING AT LEAST ONE SIGNIFICANT RESULT

Eric Derobert
Sanofi
Eric.Derobert@sanofi.com

Julie Perez

In the area of pharmaceutical statistics, the problem of multiplicity in clinical trials commonly arises. As soon as, for example, many-to-one comparisons of treatments with a control, or multiple endpoints (typically primary and secondary), are considered, solutions have to be found to control the Type I error. Among others, gatekeeping procedures are a well-known class of procedures – especially the Bretz et al. graphical approach [2] – which permit to control multiplicity testing hierarchically ordered hypotheses.

However, these gatekeeping procedures are not so often used. They are suspected for keeping power for the secondary endpoint to the detriment of results on the primary endpoint, which seems obviously irrelevant. That is the reason why we chose to focus on developing a parameterized gatekeeping strategy, based on the preliminary condition that results obtained on the secondary endpoint would not be to the detriment of proving the efficacy of at least one treatment on the primary endpoint. This work was developed in the simple case (although already computationally complex) of the comparison of two treatments versus a control with a primary and a secondary endpoint (i.e. four null hypotheses have to be tested), where Dunnett [3] and weighted Simes tests [1,4,5] are combined

Therefore, we defined a gatekeeping parameter and worked on its optimization, using a logistic regression model depending on parameters of the clinical trials such as randomization ratio, correlation between endpoints, treatment effect-sizes, risk levels. We also defined a subjective priority ratio, illustrating the relative importance given to the rejection of hypotheses for one treatment on both endpoints or of both treatments on the primary endpoint only. This subjective priority ratio turns out to be one of the most important parameters for the choice of the gatekeeping strategy in the several logistic models we suggested

The regression models we obtained permit to define an easy-to-use gatekeeping strategy, based on the values of the different parameters involved in the clinical trial. Based on these models, one major result about gatekeeping seems to be that the gatekeeping strategy is extremely sensitive to any small variation of some parameters such as effect-sizes or priority ratio

Finally, after studying the four-hypothesis gatekeeping strategy, we tried to think about possible extensions to trials testing more than two treatments on one primary and one secondary endpoints. Such extensions are not so easy to develop, mainly because of the increasing level of complexity of multinormal probability computation.

References:

[1] Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. Scandinavian Journal of Statistics, 24:407-418.

[2] Bretz, F., Maurer, W., Brannath, W. and Posch M (2009). A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine, 28:586-604.
[3] Dunnett, C. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50:1096-1121.
[4] Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. Biometrika, 73:751-754.
[5] Tamhane, A.C. and Liu, L. (2008). On weighted Hochberg procedures. Biometrika, 95:279-294.

SESSION 5 (5), JULY 6, 10:30 – 12:15

MULTIPLICITY ADJUSTMENT FOR COMPOSITE BINARY ENDPOINTS

Geraldine Rauch
University of Heidelberg
rauch@imbi.uni-heidelberg.de

Meinhard Kieser
University of Heidelberg

Binary composite outcome measures are increasingly used as primary endpoints in clinical trials. Composite endpoints combine several events of interest within a single variable. However, as the effect observed for the composite does not necessarily reflect the effects for the individual components, it is recommended in the literature to additionally evaluate each component separately. The task is to define an adequate multiple test procedure which focuses on the composite outcome measure but allows for a confirmatory interpretation of the components in case of large effects. We determine the correlation matrix for a multiple binary endpoint problem of a composite endpoint and components based on the normal approximation test statistic for rates. Thereby, we assume multinomial distributed components. We use this correlation to calculate the adjusted local significance levels. We discuss how to use our approach for a more informative formulation of the test problem. Our work is illustrated by two clinical trial examples. In conclusion, by incorporating the correlation under the null hypotheses, the global power for the multiple test problem assessing both the composite and its components can be increased as compared to simple Bonferroni-adjustment. Thus, a confirmatory analysis of the composite and its components might be possible without a large increase in sample size as compared to a single endpoint problem formulated exclusively for the composite.

SESSION 6 (1), JULY 6, 12:30 – 14:15

DESIGNING ISSUES IN POPULATION ENRICHMENT DESIGNS

Gernot Wassmer
Aptiv Solutions
gernot.wassmer@aptivsolutions.com

Silke Jürgens
Aptiv Solutions

According to Temple (1994), enrichment designs are applicable where studies of unselected patients might be unable to detect a drug effect and it seems necessary to "enrich" the study with potential responders. Using the combination testing principle together with the closed testing procedure, the definition of a seamless enrichment strategy that controls the FWE in a strong sense is straightforward (e.g., Brannath et al., 2009). It can be used for continuous, binary and survival endpoint. The methodology is implemented in the new PE module of ADDPLAN. We present designing issues that determine the statistical performance of such designs, and illustrate by examples how simulations results might help to select an appropriate design.

References:

[1] Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., Racine-Poon, A. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy on oncology. Statistics in Medicine, 28: 1445-1463.
[2] Temple, R. (1994). Special study designs: Early escape, enrichment, studies in non-responders. Communications in Statistics - Theory and Methods, 23: 499-531.

SESSION 6 (2), JULY 6, 12:30 – 14:15

CLINICAL TRIAL DESIGNS WITH DELAYED SELECTION OF THE PRIMARY COMPARISON

Ekkehard Glimm
Novartis Pharma, Basel
ekkehard.glimm@novartis.com

Recently, statisticians in the pharmaceutical industry are facing a need for more complex confirmatory trial designs. One of the drivers of this development is the improvement of diagnostic predictors (like genetic biomarkers). For example, laboratory experiments may hint at a larger treatment benefit in patients who express a certain gene. However, at the start of the clinical trial, such an increased subpopulation benefit is often still hypothetical. Hence, the confirmatory clinical trial begins with the multiple aim of (i) establishing the treatment effect in the full population, or (ii) in the subpopulation and (iii) finding out if the hypothesis about an enhanced effect in the subpopulation is true.

From the design perspective, this situation calls for designs where an interim analysis is used to decide about the primary comparison (subpopulation or full population) and potential changes in recruitment (e.g. an increase of the number of subpopulation patient in an 'enrichment design'). With respect to data analysis, the multiplicity issue arising from the comparison of treatment effects in two (sub- and full population) or three (sub-, non-sub- and full population) needs to be addressed.

In this talk, designs and analyses for such clinical trials will be discussed. The discussion also addresses other situations where similar statistical challenges arise (e.g. multiregional confirmatory trials that have to be submitted to several health authorities who are primarily interested in 'their' regional subpopulation; trials where treatments have different modes of application, each with a corresponding control treatment).

SESSION 6 (3), JULY 6, 12:30 – 14:15

MULTI-STAGE DROP-THE-LOSER DESIGNS

James Wason
MRC Biostatistics Unit, Cambridge
james.wason@mrc-bsu.cam.ac.uk

Jack Bowden
MRC Biostatistics Unit, Cambridge

A research topic of great current interest is designing multi-arm multi-stage (MAMS) trials. MAMS trials improve the efficiency of the drug development process when multiple new treatments are available for testing. A group-sequential approach can be used in order to design MAMS trials, using an extension to the Dunnett multiple testing procedure [1,2]. The expected sample size of group-sequential MAMS trials is generally low, however the actual sample size used is a random variable, which can take large values. This can often cause problems with applying for funding to conduct such a trial as an investigator would have to request sufficient funding for the maximum plausible sample size. This motivates a type of design, which provides the efficiency advantages of a group-sequential MAMS design, but has a fixed sample size. One such design is the two-stage drop-the-loser design [3], in which a number of experimental treatments, and a control treatment, are assessed at an interim analysis. The best performing experimental treatment and the control treatment then continue to a second stage. I will discuss extending this design to more than two stages, which can noticeably reduce the sample size required. I also compare the resulting sample size requirements to the sample size distribution of analogous group-sequential MAMS designs. The sample size required for a multi-stage drop-the-loser design is usually above the expected sample size of a group-sequential MAMS trial, but not by much. In many practical scenarios, the disadvantage of a slight loss in efficiency would be overcome by the huge advantage of a fixed sample size.

References:

[1] Dunnett, C. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50: 1096-1121.
[2] Magirr, D. Jaki, T. and Whitehead J. (2012). A generalized Dunnett test for multiarm multistage clinical studies with treatment selection. Biometrika (Epub).
[3] Sampson, A. and Sill, M. (2005). Drop-the-losers design: normal case. Biometrical Journal, 47: 257-268.

SESSION 6 (4), JULY 6, 12:30 – 14:15

CONDITIONALLY UNBIASED AND NEAR UNBIASED ESTIMATION FOR MULTI-STAGE DROP-THE-LOSERS DESIGNS

Jack Bowden
MRC Biostatistics Unit
jack.bowden@mrc-bsu.cam.ac.uk

Ekkehard Glimm
Novartis Pharma AG

A two-stage drop-the-losers trial provides the framework for identifying a promising experimental treatment from a group of candidates. At the mid-trial interim analysis, only the best performing treatment is selected for further study. This design has been extensively researched, see for example [1,2]. The multi-stage drop-the-losers design is a natural extension to the original idea, it enables selection to occur after several interim looks. This can markedly increase the probability of selecting the truly best treatment compared with the two-stage design, especially when the number of treatments is large. Wason and Bowden [3] have investigated hypothesis testing for this new trial scenario.

Our focus here is on estimation; building on the work of Cohen and Sackrowitz [1] and Bowden and Glimm [3], we derive the uniform minimum variance conditionally unbiased estimate (UMVCUE) of the selected treatment for the general multi-stage setting. A different derivation to that of [1] and [3] is used, based on multivariate transformations. We show that allowing additional stages of drop-the-losers selection requires an increasingly strong and unexpected form of conditioning to be employed. This motivated the suggestion of an alternative near unbiased estimator with weaker (and strictly incorrect) conditioning. We call this the 'ad-hoc UMVCUE'.

In order to best elucidate and motivate the new approach, we focus on a specific example of a three-stage drop-the-losers trial. Since each extra interim analysis leads to an increased administrative burden, it is arguably the most pertinent alternative to the original two-stage design. We compare the UMVCUE's performance against the ad-hoc UMVCUE and the bias-adjusted MLE [5] in terms of bias, mean squared error, confidence interval width and coverage. The ad-hoc UMVCUE is shown to be the most attractive all-round estimator.

References:

[1] Cohen, A, Sackrowitz (1989). Two stage conditionally unbiased estimators of the selected mean. Statistics and Probability Letters 8: 273-278.
[2] Sampson, A. and Sill, M. (2005). Drop-the-losers design: normal case. Biometrical Journal, 47: 257-268.
[3] Wason, J. and Bowden, J. (2012) Multi-stage drop-the-loser designs. Technical report, MRC Biostatistics unit, Cambridge.
[4] Bowden, J. and Glimm, E. (2008). Unbiased estimation of selected treatment means in two-stage trials. Biometrical Journal 50, 515-527.
[5] Bebu, I., Luta, G. and Dragalin., V. (2010). Likelihood inference for a two-stage design with treatment selection. Biometrical Journal 52, 811-822.

SESSION 6 (5), JULY 6, 12:30 – 14:15

FLEXIBLE TRIAL DESIGN IN PRACTICE – STOPPING ARMS FOR LACK-OF-BENEFIT AND ADDING RESEARCH ARMS MID-TRIAL IN STAMPEDE. A MULTI ARM MULTI STAGE RANDOMISED CONTROLLED TRIAL

Matthew R Sydes
MRC Clinical Trials Unit, London
matthew.sydes@ctu.mrc.ac.uk

Mahesh KB Parmar [1], Malcolm D Mason [2], Noel W Clarke [3], Claire Amos [1], John Anderson [4], Johann de Bono [5], David P Dearnaley [5], John Dwyer [6], Charlene Griffith [1], Gordana Jovic [1], Alastair Ritchiem [1], J Martin Russell [7], Karen Sanders [1], George Thalmann [8], Nicholas D James [9]
[1] MRC Clinical Trials Unit, London
[2] School of Medicine, Cardiff University, Cardiff
[3] The Christie and Salford Royal Hospitals Foundations Trusts, Manchester
[4] The Royal Hallamshire Hospital, Sheffield
[5] Institute of Cancer Research and Royal Marsden Hospitals Foundation Trust, Sutton
[6] Prostate Cancer Support Federation, Stockport
[7] Beatson West of Scotland Cancer Centre, Glasgow
[8] Inselspital, Bern
[9] School of Cancer Sciences, University of Birmingham, Birmingham

on behalf of the STAMPEDE Investigators

OBJECTIVES

STAMPEDE is a randomised controlled trial designed a novel multi arm, multi stage (MAMS) design. Here we describe methodological and practical issues arising with: (1) stopping recruitment to research arms following a pre-planned intermediate analysis, and (2) adding a new research arm during the trial.

METHODS

STAMPEDE recruits men with locally advanced or metastatic prostate cancer starting standard long-term hormone therapy. There were originally 5 research and 1 control arms, each undergoing a pilot stage (safety and feasibility), 3 intermediate 'activity' stages (I III) focusing on failure free survival (FFS), then a final 'efficacy' stage (IV) focusing on overall survival. Each research arm is formally compared in a pairwise manner to the control arm at the end of each stage. Accrual of further patients continues to the control arm and those research arms showing activity and an acceptable safety profile. At each stage, the stop ping guideline compares the observed treatment effect against a pre-defined cut off value, which becomes increasingly stringent stage by stage. The design facilitates adding new research arms should sufficiently interesting agents emerge. These are compared only to contemporaneously recruited control arm patients using the same intermediate guidelines in a time-delayed manner. The addition of new research arms is not dependent on the original research arms stopping accrual early but is subject to adequate recruitment to support the overall trial aims.

RESULTS

(1) Stopping Existing Therapy: After the second intermediate activity analysis, recruitment discontinued to two research arms for lack of sufficient activity. Detailed preparations meant that changes were implemented swiftly at 100 international centre and recruitment continued seamlessly into Activity Stage III, with 3 remaining research arms and the control arm. Further regulatory and ethical approvals were not required because this was already included in the initial trial design.

(2) Adding New Therapy: An application to add a new research arm was approved by funder, (who also organised peer review), industrial partner and regulatory and ethical bodies. This was all done in advance of any decision to stop current therapies.

CONCLUSIONS

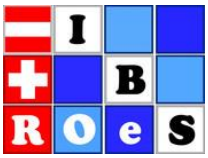
The STAMPEDE experience shows that recruitment to MAMS trial is achievable and that mid flow changes to trial design with good planning. This benefits patients and the scientific community as research treatments are evaluated more efficiently and cost effectively.

TRIAL REGISTRATION

ISRCTN78818544, NCT00268476

Sponsors


Internationale
Biometrische
Gesellschaft
Deutsche Region


I
B
R O e S


Abbott
A Promise for Life


ACCOVION


AptivSolutions
Accelerating the Possibilities


DATAMAP


MEDA


METRONOMIA
Clinical Research Services


NOVARTIS


SCHWABE
From Nature.
For Health.
Dr. Willmar Schwabe Pharmaceuticals
www.schwabepharma.com

ACKNOWLEDGEMENT

We explicitly thank the “Internationales Wissenschaftsforum Heidelberg (IWH)” for hosting this workshop.

NOTES

NOTES

NOTES