The 7th International Conference on

Multiple Comparison Procedures

August 29 – September 01, 2011

Washington D.C., USA

Imprint:

MCP 2011, Washington D.C., USA
7th International Conference on Multiple Comparison Procedures
August 29 – September 01, 2011

Production: Sophie Frantal, Elisabeth Pernicka

Table of Contents

Sponsors of the MCP 2011 and the "Society for the Support of the International MCP Conference"

The Society for the Support of the International MCP Conference wishes to acknowledge the contribution of, and to express their warm appreciation to

Accovion
Aptive Solutions
Bayer-Schering
Celgene
Cytel
Johnson & Johnson
Novartis
Quintiles
SAS
Tessella
Vertex

Keynote Speakers

- James O. Berger (Duke University, USA)
- Terry Speed (WEHI, Australia & UC Berkeley, USA)

Invited Speakers

- Peter Bauer (Medical University of Vienna)
- Andrew Gelman (Columbia University)
- Jelle Goeman (Leiden University)
- Olivier Guilbaud (AstraZeneca)
- Franz Koenig (Medical University of Vienna)
- Peter Mueller (University of Texas Austin)
- Steve Ruberg (Eli Lilly)
- Sanat Sarkar (Temple University)
- Klaus Strassburger (University of Duesseldorf)
- Michael Wolf (University of Zurich)

General Information

Conference Venue:
Hilton Hotel
1750 Rockville Pike
Rockville, Maryland
USA 20852-1699
Tel: +1-301-468-1100
Fax: +1-301-468-0308

Conference Website:
www.mcp-conference.org

Important Dates

Aug 22	End online registration deadline
Aug 29	Short courses (8:30 – 12:30 / 1:30 – 5:30)
Aug 29	Social Mixer (5:30 – 7:30)
Aug 30	Start main conference
Aug 30	Dinner banquet at conference hotel; group picture (6:30 – 9:30)
Aug 31	Conference excursion: boat cruise ride along the Potomac River
Sep 01	Conference end

Scientific Program

Short Courses

Monday, 29 August: 8:30 am – 12:30

Gatekeeping procedures in clinical trials
Alex Dmitrienko and Ajit Tamhane

Monday, 29 August: 8:30 am – 12:30

Multiple comparisons in complex clinical trial designs
H.M. James Hung and Sue-Jane Wang

Monday, 29 August: 8:30 am – 12:30

Concepts and techniques of multiple testing
in clinical and biomarker studies
Jason C. Hsu

Monday, 29 August: 1:30 -5:30 pm

Graphical approaches to multiple test problems
Frank Bretz, Ekkehard Glimm, and Willi Maurer

Monday, 29 August: 1:30 -5:30 pm

Simultaneous confidence bands in regression
Wei Liu

Monday, 29 August: 1:30 -5:30 pm

Adaptive designs for clinical trials
Martin Posch and Franz Koenig

Sessions

Tuesday, 30 August, 8:00 – 10:00 am
Keynote Session
Plaza Ballroom

Opening of the Conference – Sue-Jane Wang

James O. Berger
Bayesian Adjustment for Multiplicity

Terry Speed
Multiple testing for generalized bump-hunting in genomics

Tuesday, 30 August, 10:30 am – 12:00 pm

Adaptive Designs Chair: James Hung	Bayesian Methods Chair: Yoav Benjamini
Plaza I+II	Plaza III
Maximum type I error rate inflation of conventional tests applied in trials with (inbalanced) sample size reassessment and treatment selection Peter Bauer	**Why we (usually) don't worry about multiple comparisons** Andrew Gelman
Testing efficacy for multiple endpoints in clinical trials that allow sample size adaptation Yi Liu	Analysis of dose-response microarray data using Bayesian Variable Selection (BVS) methods: Modeling and multiplicity adjustments Ziv Shkedy
Adaptations without unblinding Martin Posch	Multiplicity-adjusted comparisons of a candidate genomic predictive model with other models in the Microarray Quality Control (MAQC)-II Study Samir Lababidi
Multiplicity in adaptive selection Design Sue-Jane Wang	Practice of using / not using multiple comparisons In medical research – ethical dilemma of a biostatistician Mathai Achirathalackal

Tuesday, 30 August, 1:30 – 3:00 pm

Adaptive Designs Chair: Franz Koenig	Decision Theory Chair: Henry Hsu
Plaza I+II	Plaza III
A graphical approach for adaptive clinical trials testing multiple hypotheses Florian Klinglmueller	**Bayesian decision theoretic MCP: application to phage display data** Peter Mueller
Confidence intervals and point estimates for adaptive group sequential trials Lingyun Liu	Binary classification with pFDR-pFNR losses Thorsten Dickhaus
Confidence Intervals for adaptive two stage designs with treatment selection Ionut Bebu	Goodness of fit, higher criticism and local levels Sandra Landwehr
Confidence intervals for adaptive two stage designs with two subpopulation Vlad Dragalin	False and accurate significance approximation for genome-wide disease association studies Yu Zhang

Tuesday, 30 August, 3:30 – 5:00 pm

Adaptive Designs Chair: Willi Maurer	Resampling based methods Chair: James Troendle
Plaza I+II	Plaza III
Use of modeling approaches to support dose selection at interim in adaptive designs for confirmatory clinical trials Franz König	Permutation multiple tests of binary features are not guaranteed to control error rates Eloise Kaizar
Partition testing for confirmatory seamless Phase II/III clinical trials with ordered doses involving multiple endpoints Toshifumi Sugitani	Permutational multiple testing adjustments with multivariate multiple group data James Troendle
A generalized Dunnett test for multiarm-multistage clinical studies with treatment selection Dominic Magirr	Resampling-based confidence regions and multiple tests Sylvain Arlot
Majesty and misery of interim dose selection (as conjectured from a 3-doses configuration) Eric Derobert	

Wednesday, 31 August, 8:30 – 10:00 am

Clinical Trials Chair: Sue-Jane Wang	Simultaneous confidence intervals Chair: Daphne Lin	Error Rates Chair: Thorsten Dickhaus
Plaza I	Plaza II	Plaza III
Panel session on a multiplicity case study Frank Bretz Alex Dmitrienko Jason Hsu James Hung Mohammad Huque Gary Koch	Simultaneous inference for the quantiles of a normal population Wei Liu	**Estimates and confidence bounds for the number of False hypotheses: A partitioning approach** Klaus Strassburger
	Confidence Bands for Distribution Functions When Parameters are Estimated from the Data: A Non Monte-Carlo Approach Walter Rosenkrantz	A sharp upper bound for the expected number of false rejections Alexander Gordon
	Multiple testing in adaptive designs Michael Rosenblum	Control of the expected number of false rejections in multiple hypotheses testing Marsel Scheer
	An interval property for multiple testing procedures Harold Sackrowitz	A sufficient and necessary condition on strong control generalized familywise error rate in multiple hypothesis testing Huajiang Li

Wednesday, 31 August, 10:30 am – 12:00 pm

Clinical Trials Chair: Sue-Jane Wang	Simultaneous confidence intervals Chair: Wei Liu	False Discovery Rate Chair: Alexander Gordon
Plaza I	Plaza II	Plaza III
Panel session on multiplicity issues Peter Bauer Willi Maurer Walt Offen Robert O'Neill Norman Stockbridge Robert Temple	**Simultaneous confidence regions for closed tests, including Hochberg's and Hommel's procedures based on p-values** Olivier Guilbaud	On the null-problem in multiple hypotheses testing Veronika Gontscharuk
	Calculation of simultaneous confidence intervals by constraint propagation Georg Gutjahr	MCP under stochastic order: controlling FDR Jinde Wang
	Multiple comparisons among components of mean vector under an elliptical population Sho Takahashi	Hierarchical testing of subsets of hypotheses Marina Bogomolov
	Conservative simultaneous confidence intervals for multiple comparisons of correlated mean vectors with a control Takahiro Nishiyama	Conservative adjustment of q-value Yinglei Lai

Wednesday, 31 August, 1:30 – 3:00 pm

Clinical Trials Chair: George Chi, Mohammad Huque	Linear Models Chair: Anna Nevius	False Discovery Rate Chair: Sanat Sarkar
Plaza I	Plaza II	Plaza III
Communicating advanced multiple test strategies to clinical teams - a case study Frank Bretz	Multiple comparisons in the ANOVA model under heteroscedasticity Gerhard Hommel	**Controlling the false discovery rate in two-stage combination tests for multiple endpoints** Sanat Sarkar
A multiple comparison procedure for hypotheses with gatekeeping structure Xiaolong Luo	Heteroscedastic analysis of means with unequal sample sizes Miin-Jye Wu	Adaptive FWER and FDR control under block dependence Wenge Guo
On validity of analysis of mortality Qing Liu	Multiple testing of composite null hypotheses in heteroscedastic models Alexander McLain	Exact calculations for the false discovery proportion and applications Etienne Roquain
Statistical consideration for the design and analysis of clinical trials with targeted subgroups Mohamed Alosh	Estimating the Largest Parameter from uniform distribution in the presence of outliers from generalized uniform distribution Sushmita Jain	Generalized stepwise procedures to control the false discovery rate Scott Roths

Thursday, 01 September, 8:30 – 10:00 am

Multiple Endpoints Chair: Yunling Xu	Gatekeeping Methods Chair: Brian Wiens	Exploratory Analysis Chair: Martin Posch
Plaza I	Plaza II	Plaza III
Designing multi-regional clinical trial with different regional required primary endpoints Yi Tsong	Development of gatekeeping procedures in confirmatory trials Alex Dmitrienko	**Cherry-picking? Multiple testing for exploratory research** Jelle Goeman
Test procedures for the assessment of the components of composite endpoints Geraldine Rauch	Multistage parallel gatekeeping with retesting George Kordzakhia	Partitioning testing for broad efficacy and efficacy in genomic subgroups Szu-Yu Tang
Statistical and regulatory consideration for multi-item patient reported outcome (PRO) Rima Izem	Reproducibility of conclusions on multiple hypotheses from one or more families Brian Wiens	Correction of the significance level after multiple coding of an explanatory variable in generalized linear model. Jérémie Riou
Establishing non-inferiority and equivalence in matched pair designs with multiple endpoints based on McNemar's test Jin Xu	Sequentially rejective graphical multiple test procedures with memory Willi Maurer	Stability based testing for the analysis of fMRI data Joke Dumez

Thursday, 01 September, 10:30 am – 12:00 pm

Multiple Endpoints Chair: Mohamed Alosh	Clinical Trials Chair: George Kordzakhia	Subgroup Analysis Chair: Alex Dmitrienko
Plaza I	Plaza II	Plaza III
Testing multiple endpoints in complex clinical trial designs James Hung	Resolving the Type I and Type II error dilemma for clinical safety analyses Devan Mehrotra	**Challenges in developing tailored therapeutics to improve personalized medicine** Steve Ruberg
An adaptive extension of a two-stage group sequential procedure for testing a primary and a secondary endpoint with gatekeeping constraint Ajit Tamhane	Analysis of multi-regional clinical trials: Applying a two-tier procedure to decision-making by individual local regulatory authorities Yunling Xu	A novel recursive partitioning method for establishing response to treatment in subpopulations Ilya Lipkovich
A nonparametric procedure to compare clustered multiple endpoints Aiyi Liu	Multiple testing with latent variable model for ordered categorical response Tong-Yu Lu	Interaction trees for subgroup analysis Xiaogang Su
Graphical approaches for multiple endpoint problems using weighted parametric tests Ekkehard Glimm	Optimizing drug development: An application to diabetes Zoran Antonijevic	Identifying subgroups in clinical trials via random forests and regression trees Jared Foster

Thursday, 01 September, 1:30 – 3:00 pm

Software Solutions Chair: Vlad Dragalin	Closed Testing With Applications Chair: Ajit Tamhane
Plaza I	Plaza II
muTOSS - Multiple hypotheses testing in an open software system Wiebke Werft	**Consonance and the closure method in multiple testing** Michael Wolf
gMCP: A graphical user interface for graphical described multiple test procedures Kornelius Rohmeyer	A consonant partition testing Strategy for Multiple Endpoints Bushi Wang
SiZ-MCP: A new tool for sample size calculations for MCPs Cyrus Mehta	Joint models and tests for time to tumor recurrence and disease stage in Oncology clinical trials Olga Marchenko
New SAS tools for multiple comparisons in very general models Randy Tobias	Alpha maximized multiplicity adjustment in genomic studies using sequential post-hoc matching Jimmy Efird

Wednesday, 31 August, 3:00 – 4:00 pm

Poster session
Combining p-values from independent studies JaiWon Choi
Bayesian testing for no effect in nonparametric regression Taeryon Choi
Sample size determination in clinical trials with two correlated co-primary time-to-event endpoints Toshimitsu Hamasaki
Multiple testing procedures with applications to whole-genome analysis Hongmei Jiang
Tests for two mean vectors and simultaneous confidence intervals with unequal covariance matrices in two-step monotone missing data Tamae Kawasaki
A tow stage procedure to control the generalized family wise error rate Djalel Eddine Meskaldji
On the distributions of some test statistics for profile analysis with two-step monotone missing data Mizuki Onozawa
Testing the equality of pairs of mean vectors and simultaneous confidence intervals in elliptical distributions Aya Shinozaki and Takashi Seo
On the identification of predictive biomarkers in high-dimensional data Wiebke Werft
Sample size calculation in Phase II selection designs Zuoshun Zhang

Talks

Abstracts are sorted by session.
The code at the bottom of each abstract denotes the session and is of the form: Day (Tu,W,Th), Session (am1, am2, pm1, pm2), Plaza (PI, PII, PIII), Talk (T1-T5).

Keynote

Bayesian Adjustment for Multiplicity

James O. Berger

Duke University, USA

Issues of multiplicity in testing are increasingly being encountered in a wide range of disciplines, as the growing complexity of data allows for consideration of a multitude of possible hypotheses; failure to properly adjust for multiplicities is possibly to blame for the apparently increasing lack of reproducibility in science. Bayesian adjustment for multiplicity is interesting, in that it occurs through the prior probabilities assigned to models/hypotheses. It is, hence, independent of the error structure of the data, the main obstacle to adjustment for multiplicity in non-Bayesian statistics.

Not all assignments of prior probabilities adjust for multiplicity, however, and assignments in huge model spaces typically require a mix of subjective assignment and appropriate hierarchical modeling. These issues will be reviewed through a variety of examples. If time permits, some surprising issues will also be discussed, such as the fact that empirical Bayesian approaches to multiplicity adjustment can be problematical.

Tuam1PBallroomT1

Keynote

Multiple Testing for Generalized Bump-Hunting in Genomics

Terry Speed

Walter & Eliza Hall Institute of Medical Research, Australia

In one dimension, the term bump-hunting has traditionally been used to denote searching for modes in probability density, rate, intensity or regression functions. With generalized bumphunting I'm extending this usage to cover searching for regions of interest which can be either troughs or peaks in a random function along the genome, relative to some reference value.

In genomics, the problems typically arise as follows. They begin with genome-wide data, either at the base pair level, as with DNA-seq, ChIP-seq or RNA-seq, or at a lower resolution, as with tiling or methylation microarrays. (I'll define any terms like this that I need in my talk.) These data may come from a single genome, a matched or unmatched pair of genomes, a single sample of genomes, or two or more samples of genomes. Some statistical or computational analysis is carried out on the data, leading to a collection of points or intervals along the reference genome, which will usually be smoothed in some way, resulting in a function along that genome. This function is then subject to further analysis, for example, by thresholding or local testing, giving rise to a collection of regions of interest, which are intervals constituting peaks or troughs of the function. Interest focuses not only on the centres of these regions (bumps), but their extent as well, that is, the entire interval, and the regions don't need to be "bump-like," they can have more or less arbitrary shapes, as long as they satisfy their defining property.

From a scientific viewpoint, we know that some, perhaps many of these generalized bumps are likely to be real, that is, to be consequences of known biological processes, which will be validated by further assays. We also expect that some, perhaps many of them will simply be random variation, or noise, that is, false positives, which would not be validated. The problem, as is to be expected, is to distinguish the true from false bumps, and do so in a principled way,
controlling the genome-wide error rate in some way. For example, could we aim for a specified false discovery rate, where each bump

found is a “discovery”? While it might appear that each bump poses a hypothesis testing problem, it is may be important to note that the hypotheses here are not, in general, specified in advance of collecting and analysing the data. In my talk I will give a couple of examples of generalized bump hunting in genomics, discuss some of what is currently done about testing in this context, and pose and attempt to answer some questions. It seems to me that this class of problems may warrant the attention of the multiple comparisons community.

Tuam1PBallroomT2

Maximum Type I Error Rate Inflation of Conventional Tests applied in Trials with (Inbalanced) Sample Size Reassessment and Treatment Selection

Peter Bauer, Alexandra Graf

Medical University of Vienna, Austria

Sample size reassessment in an adaptive interim analysis based on an estimate of the effects size can considerably increase of the type I error rate if the pre-planned conventional fixed sample size test is applied in the final analysis. For the comparison of the means of two independent normal distributions we first extend known results to the case when the total sample size and the allocation rate to the two treatment arms can be modified in the interim analysis. Then we consider the case that more than one treatment is compared to a control and treatment selection is performed at interim, e.g., going on with the “best” treatment and the control and dropping all other treatments at interim. The maximum inflation of the type I error rate in this many-one multiple comparisons scenario can be calculated by searching for the “worst case” adaptation rules which, given the interim data, lead to the largest conditional type error rate of the conventional fixed sample size test in any point of the interim sample space. When allocation ratios are modified the calculation of the conditional type I error rate requires the knowledge of a nuisance parameter, the common mean under the global null hypothesis. In practice this assumption may apply at least to a good approximation when a standard control treatment is used for which precise estimates are available from extensive historical data. The maximum inflations of the type I error rate may become substantially larger than that derived by Proschan and Hunsberger (1995) for sample size reassessment balanced between treatments.

Proschan, M.A. and Hunsberger, S.S. (1995). Designed extensions of studies based on conditional power. Biometrics 51, 1315 -1324.

Graf, A.C. and Bauer, P. (2011). Maximum inflation of the type I error rate when sample size and allocation rate are adapted in a pre-planned interim look. Statistics in Medicine, electronic preview.

Tuam2PI+IIT1

Testing Efficacy for Multiple Endpoints in Clinical Trials that allow Sample Size Adaptation

Yi Liu, Mingxiu Hu, Hua Liu, Hongliang Shi

Millennium: The Takeda Oncology Company, USA

A lot of sample size re-estimation methods have been proposed and discussed in the recent literature. Implemented with care, these designs can gain power over the traditional fixed sample size or the group sequential design while still preserving the Type I error rate. However, most of them are restricted to testing a single hypothesis. In this paper, we propose designs that apply certain sample size adaptation rules at the interim analysis in a group sequential setting that aims to claim efficacy for multiple endpoints. A theoretical proof of the strong control of overall Familywise Error Rate (FWER) is provided. Power comparisons with the traditional group sequential method with multiple endpoints are performed and discussed.

Tuam2PI+IIT2

Adaptations without Unblinding

Martin Posch, Michael Proschan

European Medicines Agency, UK

Regulatory guidelines favour blinded over unblinded interim analysis wherever feasible. This talk concerns conditions for a valid analysis when an adaptation is made before unblinding. In the context of non-parametric as well as parametric testing procedures, we give examples where adaptations based on blinded data preserve the validity of hypotheses tests and explore settings where they may lead to bias.

Tuam2PI+IIT3

Multiplicity in Adaptive Selection Design

Sue-Jane Wang

U.S. Food and Drug Administration, USA

It is well recognized that adaptive design allows multiple ways to win due to the choices among the multiple hypotheses initially set out with. The FDA draft guidance on adaptive design lays out some pivotal foundations for adaptive designs to be properly considered in clinical development programs. In light of the draft guidance, the recent advances on clinical trial methodology provide an opportunity to take a fresh look of fixed designs, group sequential designs and a broader class of adaptive designs. In regulatory submissions, adaptive design has been used in early phase and late phase clinical trials. The implication of an adaptive design proposal depends on the stage in a drug development program. In this presentation, I shall use methodological approaches to discuss the probability of correct selection and the control of type I error rate due to multiplicity with an adaptive approach. I shall also revisit the thought process necessary to entertain usage of adaptive designs for dose regimen and/or patient population adaptive selection in the clinical development program, distinguishing between learning stage and confirmatory stage.

Tuam2PI+IIT4

Why we (usually) don't worry about Multiple Comparisons

Andrew Gelman, Jennifer Hill, Yu-Sung Su

Columbia University, USA

Applied researchers often find themselves making statistical inferences in settings that would seem to require multiple comparisons adjustments. We challenge the Type I error paradigm that underlies these corrections. Moreover we posit that the problem of multiple comparisons can disappear entirely when viewed from a hierarchical Bayesian perspective. We propose building multilevel models in the settings where multiple comparisons arise.

Multilevel models perform partial pooling (shifting estimates toward each other), whereas classical procedures typically keep the centers of intervals stationary, adjusting for multiple comparisons by making the intervals wider (or, equivalently, adjusting the p-values corresponding to intervals of fixed width). Thus, multilevel models address the multiple comparisons problem and also yield more efficient estimates, especially in settings with low group-level variation, which is where multiple comparisons are a particular concern.

Tuam2PIIIT1

Analysis of Dose-Response Microarray Data using Bayesian Variable Selection (BVS) Methods: Modeling and Multiplicity Adjustments

Ziv Shkedy, Adetayo Kasim, Dan Lin

Hasselt University, Belgium

The biotechnology of DNA microarrays allow the monitoring expression levels of thousands of genes simultaneously, and identifying those genes that are differentially expressed. As a result type I error (the probability for false identification) increase sharply when the number of tested genes gets large.

In this chapter we focus on hierarchical Bayesian modeling of dose response microarray data from early drug development experiments. We focus on a dose-response microarray experiments with four dose levels (3 microarrays at each dose level). We formulate an order restricted hierarchical Bayesian model for dose-response data and presents examples to illustrate the estimation procedures.

Within the hierarchical Bayesian framework, one of the major challenges is related to the question how to perform Bayesian inference and in particular how to adjust for multiplicity. We discuss the Bayesian Variable Selection (BVS) method which we use in order to calculate the posterior probability of a specific model given the data and the model parameters. Following Newton (2004), we use the posterior probability of the null model (of no dose effect) in order to control for multiplicity using the direct posterior probability for multiplicity adjustment.

The proposed method is applied to a dose-response experiment with 12 samples and 16998 genes.

Tuam2PIIIT2

Multiplicity-Adjusted Comparisons of a Candidate Genomic Predictive Model with other Models in the Microarray Quality Control (MAQC)-II Study

Samir Lababidi, Gene Pennello, Rong Tang

U.S. Food and Drug Administration, USA

The Microarray Quality Control Project, Phase II, (MAQC-II) was a large international effort investigating common practices to develop and validate predictive models from microarray data. Thirty-six teams built 18,303 models to predict one of 13 binary endpoints from three toxicological and three clinical training datasets. The models were then evaluated on independent validation datasets for predictive accuracy. For each endpoint, a committee selected a candidate model based on its cross-validation performance in the training set as well as the reproducibility and robustness of its development plan. In this paper we compare the candidate model with a select number of other models based on their performance in independent datasets, using frequentist and Bayesian multiple comparison procedures. A challenge is to develop a procedure that provides control on the number falsely significant results, yet achieves power by exploiting correlations among binary model predictions. Despite being a common task, the problem of comparing multiple classifiers on the same dataset has received relatively little attention in the literature. In the talk, we will discuss our two multiple comparison procedures, their findings in the MAQC study, and process by which the MAQC consortium selected the candidate model.

Tuam2PIIIT3

Practice of using / not using Multiple Comparisons in Medical Research – Ethical Dilemma of a Biostatistician

Mathai Achirathalackal

MakroCare Clinical Research Limited, India

As a Biostatistician there is various occasions in which, we are in dilemma whether to use multiple comparison procedures or not. Many occasions, the Biostatistician's consciousness says that the multiple comparison procedure should be applied. However, the circumstances compel him to be silent especially, when the customer is a postgraduate student. For the completion his / her course, he has to complete a small thesis work with fancy number of P<0.05. Then the student as well as the guide / supervisor would be happy and thesis work would be labeled as fantastic. The author would like to share his experience of being a Biostatistician in the medical field for more than 20 years. He experienced a lot of dilemma whether to go ahead with multiple comparisons or not. This is a question of Statistical ethics, but on the other hand, he needs to be considered the cry of the student who is begging for a significant result with a magic p-value of <0.05. The Biostatistician should deal this situation in a diplomatic way to apply his statistical ethics and satisfy the customer. The author would like to share his experience by presenting case studies and it would be discussed.

Tuam2PIIIT4

A Graphical Approach for Adaptive Clinical Trials testing Multiple Hypotheses

Florian Klinglmueller, Franz Koenig, Martin Posch

Medical University of Vienna, Austria

The graphical approach [1] provides a convenient tool for the definition of closed testing procedures based on weighted tests or each intersection hypothesis. We adopt this approach to construct adaptive tests for clinical trials with an unblinded interim analysis that reflect the complex contextual relations between multiple hypotheses in clinical trials. Adaptive designs are an attractive choice for confirmatory clinical trials as they provide type I error control while permitting certain mid-trial design modifications based on internal and external information, e.g, changing the pre-planned sample size, inserting/dropping of treatment groups and endpoints in clinical trials. The discussed approach is based on the closed testing principle combined with the conditional error principle. Starting with a closed testing procedure based on weighted Bonferroni tests we construct, for all intersection hypotheses, a second stage test at levels equal or smaller than the sum of marginal conditional error levels of the initial tests [2,3]. In contrast to other methods [4] the knowledge of the multivariate distribution of the test statistics is not required when using marginal conditional errors making the proposed approach, suitable for, e.g. comparing treatment groups and/or multiple endpoints.

[1] Bretz F, Maurer W, Brannat W, Posch M, (2008) A graphical approach to sequentially rejective multiple testing procedures. Stat Med 28/4, 586-604
[2] Posch M, Futschik A (2008) A Uniform Improvement of Bonferroni-Type Tests by Sequential Tests JASA 103/481, 299-308
[3] Posch M, Maurer W, Bretz F (2010) Type I error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim Pharm Stat
[4] Koenig F, Brannath W, Bretz F, Posch M, (2008) Adaptive Dunnett Tests for Treatment Selection Stat Med 27/10, 1612-25

Tupm1PI+IIT1

Confidence Intervals and Point Estimates for Adaptive Group Sequential Trials

Lingyun Liu, Cyrus Mehta, Ping Gao,Pralay Senchaudhuri, Pranab Ghosh

Cytel Inc., USA

Adaptive sequential designs have been intensively investigated in the literature. It is well known that type I error can be preserved by preserving the conditional type I error. The inference problem was addressed by Mehta et al (2007). This approach (RCI), however, is only guaranteed to provide conservative coverage of the treatment effect. In addition, this method cannot produce an unbiased point estimate. Brannath et al (2009) generalizes the stage wise adjusted confidence intervals (SWACI) of Tsiatis et al (1984) to adaptive setting. This method provides nearly exact coverage. Both of these two methods are implemented in East®.

The SWACI method is limited to one-sided test and is only applicable when there is a single adaptive change through the whole trial. For one-sided test, the SWACI method can only provide either lower or upper confidence limits but not both at the same time. We offer another approach which provides exact coverage and can be applied to a trial with multiple adaptive changes. Both confidence limits can be obtained using this new approach.

Tupm1PI+IIT2

Confidence Intervals for Adaptive Two Stage Designs with Treatment Selection

Ionut Bebu, Vladimir Dragalin, George Luta

Georgetown University, UK

The construction of adequate confidence intervals for adaptive two stage designs remains an area of ongoing research. Despite their relative simplicity, the conditional likelihood confidence intervals for the two treatments case proposed by Bebu, Luta and Dragalin (2010) compare favorably with alternative methods. First, we extend those methods to the case of more than two treatments while using higher order inference methodology. A small simulation study is used to evaluate the performance of the new methods. Second, we investigate other extensions of practical interest and illustrate them using real data, including the selection of more than one treatment for the second stage, selection rules based on both efficacy and safety endpoints, the inclusion of a control/placebo arm, covariate adjustment, and the binomial case. Although conceptually simple the new methods have a wider scope than the methods currently available.

1. Bebu, I., Luta, G. and Dragalin, V. (2010). Likelihood Inference for a Two-stage Design with Treatment Selection, Biometrical Journal, 52, 811–-822.

Tupm1PI+IIT3

Confidence Intervals for Adaptive Two Stage Designs with Two Subpopulations

Vladimir Dragalin, Ionut Bebu

ADDPLAN, An Aptiv Solutions company, USA

We consider a two-stage design of phase III randomized clinical trials for the evaluation of a tailored therapy when there is an assay predictive of which patients will be more responsive to the experimental treatment than to the control regimen. The trial starts enrolling all patients. A prospectively planned interim analysis on the primary efficacy endpoint is performed at the end of Stage I and a decision is made on whether to maintain the original study plan of showing superiority in the full population or to modify the plan to recruit only the patients classified as predictor positive in Stage II and to test superiority only in this subpopulation at the end of the study. Although there are several approaches that adjust for such adaptation and strongly control the type I error rate, no results exist on point estimation of and confidence intervals for the treatment effect at the end of such an adaptive trial. We propose such estimators and confidence intervals based on conditional likelihood method and compare them with the naive ones that ignore the fact that an adaptation has been preplanned. Relative efficiency depends upon the distribution of treatment effect across patient subsets, prevalence of the subset of patients who respond preferentially to the tailored therapy, and the assay positive predictive value.

Tupm1PI+IIT4

Bayesian Decision Theoretic MCP: Application to Phage Display Data

Peter Mueller, Luis Leon, Kim-Ahn Do

University of Texas at Austin, USA

We discuss inference for a phage display experiment with three stages. The data are tri-peptide counts by organ and stage. The primary aim of the experiment is to identify ligands that bind with high affinity to a given organ. We formalize the research question as inference about the monotonicity of mean counts over stages. The inference goal is then to identify a list of peptide and organ pairs with significant increase over stages. The desired inference summary as a list peptide and organ pairs with signficant increase involves a massive multiplicity problem. We consider two alternative approaches to address this multiplicity issue. First we propose an approach based on the control of the posterior expected false discovery rate. We notice that the implied solution ignores the relative size of the increase. This motivates a second approach based on a utility function that includes explicit weights for the size of the increase.

Tupm1PIIIT1

Binary Classification with pFDR-pFNR Losses

Thorsten Dickhaus,

Humboldt University Berlin, Germany

We draw a connection between multiple testing and binary classification and derive a false discovery rate-based approach towards the binary classification problem generalizing similar ideas in Storey (2003) to the scope of weak (two-class) mixture models. Main assets of the resulting classification algorithm are that it allows incorporating prior knowledge about class probabilities and user-supplied weighting of the severity of classification errors in both directions. The key mathematical tools to be employed are multivariate estimation methods for densities and/or likelihood ratios.
The approach was inspired and its practicability will be demonstrated by applications from the field of brain-computer interfacing and the processing of electroencephalography data.

Storey, John D. (2003). The positive false discovery rate:
A Bayesian interpretation and the *q*-value. Ann. Stat. 31, 6, 2013-2035.

Tupm1PIIIT2

Goodness of Fit, Higher Criticism and Local Levels

Sandra Landwehr, Helmut Finner, Veronika Gontscharuk

Heinrich Heine University, Germany

When testing *n* hypotheses, p-values under null hypotheses are often uniformly distributed on [0,1]. Assuming their independence, the corresponding empirical distribution function of all p-values under the global null hypothesis converges to the diagonal as n increases. Hence, under independence, one may look at a multiple test as a goodness of fit test for uniformity. One of the most well-known tests of fit is the classical Kolmogorov-Smirnov test. This test has low power against alternatives which primarily deviate from the null in the tails.

However, alternatives of this kind are indeed common. For instance, in genome-wide studies on gene expressions we face a huge number of hypotheses from which often a relatively small amount is non-null. In these situations we expect that only small order statistics of p-values may originate from non-null distributions. Then, another well-known goodness of fit test, namely the Higher Criticism introduced by Tukey [1], seems to be a more promising approach. For example, Eicker [2] has shown that the Higher Criticism is asymptotically sensitive for certain intermediate order statistics. Donoho and Jin [3] have even proved that a version of Higher Criticism is successful in the same situations where the likelihood ratio test would succeed. In general, for a multiple test procedure it is an interesting issue to consider for each order statistic of p-values the chance to exceed the corresponding critical value. We call these probabilites local levels. In this talk we present results related to the study of local levels of Kolmogorov-Smirnov and Higher Criticism giving some insight into the relationship between multiple test procedures and goodness of fit tests.

[1] Tukey, J.W. (1976), T13 N: The higher criticism. Course Notes, Statistics 411, Princeton Univ.

[2] Eicker, F. (1979), The asymptotic distribution of the suprema of the standardized empirical processes. Ann. Stat. 7, 116 – 138.

[3] Donoho, D. and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures. Ann. Stat. 32, 962 - 994.

Tupm1PIIIT3

False and Accurate Significance Approximation for Genome-Wide Disease Association Studies

Yu Zhang, Jun S Liu

Penn State University, USA

Genome-wide association studies (GWAS) commonly involve simultaneous tests of millions of single nucleotide polymorphisms (SNP) for disease association. The SNPs in nearby genomic regions are highly correlated due to linkage disequilibrium (LD, a genetic term for correlation). Simple Bonfferroni correction for multiple comparisons is too conservative. Permutation tests are on the other hand computationally expensive and are limited in scopes. We present an accurate and computationally efficient method, based on Poisson heuristic, to approximate genome-wide significance of SNP associations. Our method accurately and robustly computes p-values adjusting for millions of correlated comparisons within seconds. We demonstrate both analytically and empirically that the accuracy and efficiency of our method are nearly independent of the sample size, the number of SNPs, and the scale of p-values to be adjusted. GWAS signals tend to be small and hard to replicate due to genetic heterogeneity, rare causal mutations, and epistasis. It is thus desirable to identify weak disease associations, for which a measure of false discovery rate (FDR) is desirable. We further discuss a new approach to define and estimate FDR in GWAS.

Tupm1PIIIT4

Use of Modeling Approaches to Support Dose Selection at Interim in Adaptive Designs for Confirmatory Clinical Trials

Franz Koenig, Frank Bretz, Bjoer Bornkamp, Alexandra Graf

Medical University of Vienna, Austria

Adaptive designs for confirmatory clinical trials have received increased attention in the past years because they offer the possibility to improve the efficiency of late phase development programs (e.g.. Koenig et al 2008, Bretz, Koenig et al 2009).

In this presentation we investigate the use of modelling approaches to (i) increase the power of declaring effective dose statistically significant, (ii) support dose selection at an interim analysis. First, for a fixed sample design we apply the MCP-mod approach from Bretz et al. (2005), who suggested calculating optimal contrasts based on a-priori information about plausible dose response shapes available at the planning stage of a clinical trial, together with the closed testing procedure from Marcus et. (1976) to obtain confirmatory p-values for the global trend assessment (i.e whether there is any statistical evidence for a dose-related drug effect) as well as for the pairwise comparison of the individual doses against placebo.

In a second step we extend this closed MCP-Mod methodology to adaptive two-stage designs by applying an adaptive combination test to each intersection hypothesis (Bauer and Kieser, 1999; Hommel 2001). Combining the data from both stages in adaptive confirmatory designs allow for flexible interim decisions based on all (interim) data available of the ongoing trial while always ensuring strict type I error control. In particular, the MCP-Mod approach can be used to obtain model-based dose effect estimates at interim to guide early futility stopping and/or re-design the second stage (e.g. choice of doses, sample size, allocation ratio) and analysis (e.g., dropping of inadequate response models).

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

Bretz F, Koenig F, Brannath W, Glimm E, Posch M (2009). Adaptive Designs for Confirmatory Clinical Trials. Statistics In Medicine 2009 Apr 15;28(8):1181-217.

Bretz, F., Pinheiro, J.C., and Branson, M. (2005) Combining multiple comparisons and modeling techniques in dose-response studies. Biometrics, 61(3), 738-748.

Hommel, G. (2001). Adaptive modidcations of hypotheses after an interim analysis. Biometrical Journal, 43(5):581-589.

Koenig F., Brannath W, Bretz F, and Posch M (2008). Adaptive Dunnett Tests for Treatment Selection. Statistics in Medicine. 10:1612-25.

Marcus R, Peritz E, Gabriel KR. On closed testing procedure with special reference to ordered analysis of variance. Biometrika 1976; 63:655–660.

Tupm2PI+IIT1

Partition Testing for Confirmatory Seamless Phase II/III Clinical Trials with Ordered Doses involving Multiple Endpoints

Toshifumi Sugitani, Chikuma Hamada

Tokyo University of Science, Japan

Most widely used for adaptive seamless phase II/III designs are adaptive combination tests, which are hybrid methods that combine techniques of closed test procedures and combination tests. They use accumulating data to decide during the conduct of the study how to modify aspects of the study without undermining the validity and integrity. Nevertheless, adaptive combination tests have a crucial disadvantage, that is, they are non-consonant closed test procedures. Hence, adaptive combination tests require the evaluation of $O(2^k)$ intersection hypotheses for testing k elementary null hypotheses. This is quite computer intensive and the resulting test procedure quickly becomes intractable in multiple primary and sencondary endpoints settings. Furthermore, adaptive combination tests are not able to incorporate hierarchy between primary and secondary endpoints into its decision process. In this talk we present a hybrid method that combines the technique of partition testing with that of combination tests for adaptive seamless phase II/III designs with ordered doses involving multiple endpoints. The proposed method is simple even in the multiple primary and secondary endpoints settings and can incorporate the hierarchy between primary and secondary endpoints while preserving multiple level alpha. The performance of our proposed method is investigated in terms of the type I error and the statistical power, comparing to adaptive combination tests, via simulation study.

Key words: Adaptive seamless phase II/III design, Multiple primary and secondary endpoints, Partition Testing, Adaptive combination test, Consonance

Tupm2PI+IIT2

A Generalized Dunnett Test for Multiarm-Multistage Clinical Studies with Treatment Selection

Dominic Magirr, Thomas Jaki, John Whitehead

Lancaster University, UK

When several experimental treatments for the same disorder become available simultaneously, there are economies of scale if those treatments can be compared with a common control in a single trial. Efficiency can be further increased by monitoring the comparisons at interim analyses. Treatments that show little promise may be dropped, allowing the remaining resources to be spent on more promising treatments. Also, the trial may be stopped early if the efficacy of a treatment is already established at interim. In this talk we describe the computation of efficacy and futility boundaries for a flexible multiarm-multistage trial. The method may be seen as a generalization of Dunnett's (1955) procedure for comparing several treatments with control in a single stage trial. It will be shown that the boundaries control the family-wise error rate in the strong sense. The method is applicable for any number of treatment arms, number of stages and number of patients per treatment per stage. It can be used for a wide variety of boundary shapes or rules derived from alpha-spending functions. Sample size can be computed from a power requirement based on a 'least favorable configuration' of treatment effects.

We apply our approach to the design of a trial comparing four treatments with control in reducing insulin resistance.

Tupm2PI+IIT3

Majesty and Misery of Interim Dose Selection (as conjectured from a 3-Doses Configuration)

Eric Derobert, Fanny Windenberger

Sanofi-Aventis, France

In inferential phase II-III seamless designs (reducing phases II and III into a single clinical trial), an Interim Analysis (IA) allows to select doses to be kept until the end of the trial. By using sequentially the information, this kind of adaptive design (also considered, in this work, for a full phase II study devoted to the choice of one or two doses in a future phase III study) is expected to be more efficient than ordinary fixed designs.

After having discretized and constrained the usual case of three doses candidates (adjacency of the doses selected at the IA; no more than one local optimum in the dose response function), the research articulates in two stages:

- identify the best multiple comparisons procedures (many-to-one comparisons vs placebo) to be used in fixed design analyses, taking into consideration the uncertainty about the true unknown dose response profile,
- combine these chosen procedures for adaptive designs (including the research of optimal futility rules and optimal time for the IA) and compare their performance with that obtained for fixed designs (comparison is made easier by working wlog with a fixed number of available patients, i.e. after the IA, all the remaining patients are allocated among the remaining groups).

A particular focus is on the comparative effect of unbalancing treatment groups in fixed and adaptive designs on the power for detecting at least an interesting dose and on a clinical score built to reflect the quality of decisions taken following the dose selection. The problem of the latency period (measured by the percentage of patients not in the IA, but recruited before knowing the results of the IA) is also considered.

Tupm2PI+IIT4

Permutation Multiple Tests of Binary Features are not guaranteed to control Error Rates

Eloise Kaizar, Yan Li, Jason C. Hsu

Ohio State University, USA

Multiple testing for significant association between predictors and responses has a wide array of applications. One such application is pharmacogenomics, where testing for association between responses and many binary genetic markers is of interest. Permuting response group labels to generate a reference distribution is often thought of as a convenient thresholding technique that automatically captures dependence in the data. In reality, non trivial model assumptions are required for permutation testing to control multiple testing error rates. When binary predictors (such as genetic markers) are individually tested by standard tests, permutation multiple testing can give incorrect unconditional and, especially, conditional assessment of significances, and thus misleading results.

Tupm2PIIIT1

Permutational Multiple Testing Adjustments with Multivariate Multiple Group Data

James Troendle, Peter Westfall

National Institutes of Health, USA

Consider the multiple comparison problem where multiple outcomes are each compared among several different collections of groups in a multiple group setting. In this case there are several different types of hypotheses, with each specifying equality of the distributions of a single outcome over a different collection of groups. Each type of hypothesis requires a different permutational approach. We show that under a certain multivariate condition it is possible to use closure over all hypotheses, although intersection hypotheses are tested using Boole's inequality in conjunction with permutation distributions in some cases. Shortcut tests are then found so that the resulting testing procedure is easily performed. The error rate and power of the new method is compared to existing competitors through simulation of correlated data. An example is analyzed, consisting of multiple adverse events in a clinical trial.

Tupm2PIIIT2

Resampling-Based Confidence Regions and Multiple Tests

Sylvain Arlot, Gilles Blanchard, Etienne Roquain

Centre national de la recherche scientifique, France

We study generalized bootstrap confidence regions for the mean of a random vector whose coordinates have an unknown dependency structure. The random vector is supposed to be either Gaussian or to have a symmetric and bounded distribution. The dimensionality of the vector can possibly be much larger than the number of observations and we focus on a non-asymptotic control of the confidence level, following ideas inspired by recent results in learning theory.

We consider two approaches, the first based on a concentration principle (valid for a large class of resampling weights) and the second on a direct resampled quantile, specifically using Rademacher weights. Several intermediate results established in the approach based on concentration principles are of self-interest. We also discuss the question of accuracy when using Monte-Carlo approximations of the resampled quantities.

We present an application of these results to the one-sided and two-sided multiple testing problem, in which we derive several resampling-based step-down procedures providing a non-asymptotic FWER control. We compare our different procedures in a simulation study, and we show that they can outperform Bonferroni's or Holm's procedures as soon as the observed vector has sufficiently correlated coordinates.

(Joint work with Gilles Blanchard and Etienne Roquain. The Annals of Statistics 38, 1 (2010) 51-99.)

Tupm2PIIIT3

Multiple Testing in Adaptive Designs

Michael Rosenblum, Mark van der Laan

Johns Hopkins Bloomberg School of Public Health, USA

We propose a general framework for adapting the sampling population at an intermediate stage in a trial, testing the null hypothesis of no treatment effect for the different sampling populations, controlling family wise-error.

Wam1PIIT3

Simultaneous Inference for the Quantiles of a Normal Population

Wei Liu, F. Bretz, A.J. Hayter, E. Glimm

University of Southampton, UK

While the mean and the variance are important features of a population, many real problems require information on certain quantiles of the population which combine both the mean and variance. To make inference about several quantiles of interest simultaneously, it is appropriate to use a set of simultaneous confidence intervals for the quantiles. In this paper, a set of exact *1-α* level simultaneous confidence intervals for several quantiles of a normally distributed population is provided based on a simple random sample. With the software available, the methodology is easy to implement and illustrated with an example.

Wam1PIIT1

Confidence Bands for Distribution Functions when Parameters are estimated from the Data: A Non Monte-Carlo Approach

Walter Rosenkrantz

University of Massachusetts Amherst, USA

A method is given for computing simultaneous confidence intervals for order statistics coming from a distribution depending on one, or more, parameters that must be estimated from the data. This produces a confidence band for the distribution itself and may be regarded as an extension of Kolmogorov's goodness-of-fit test to the case where the distribution depends on parameters that must be estimated from the data. The method works whenever the joint confidence set for the parameters is convex and the quantile function is linear in the parameters. Two important special cases are treated in some detail: the normal and exponential distributions. Graphical representations and comparisons with results obtained by Lillifors and Stephens via Monte-Carlo methods are discussed. This idea of exploiting convexity to obtain simultaneous confidence intervals for linear functions of a vector parameter undoubtedly goes back to Scheff'e who derives the S-method of simultaneous multiple comparisons by exploiting the convexity of a confidence ellipsoid. An unusual feature of this paper is that we found it necessary to first prove that the joint confidence set for the mean and variance for the normal distribution based on the Wald statistic is convex and compact.

Our proof relies on an elementary theorem from differential geometry in the large due to H. Hopf and is of independent interest.

Key Words: Normal distribution function, Wald statistic, Convex confidence sets, H. Hopf's characterization of convexity, Empirical distribution function, Empirical quantile function, Kolmogorov-Smirnov, simultaneous confidence intervals.

Wam1PIIT2

Confidence Bands for Polynomial Regressions based on the Volume-of-Tube Method

Satoshi Kuriki, Naohiro Kato

The Institute of Statistical Mathematics, Japan

We provide simultaneous confidence bands for polynomial regressions when the explanatory variable is restricted to an interval [a,b], say. (a and/or b may be infinite.) According to the volume-of-tube method, the confidence bands can be constructed by evaluating the geometric invariants of the nonnegative polynomial cone (the cone consisting of polynomials that is nonnegative over the interval [a,b]) and its dual (the moment cone). Thanks to the representations of the nonnegative polynomial cone and its dual by Karlin and Studden (1966), we provide simultaneous confidence bands for polynomial regressions with the degree up to 4. The corresponding likelihood ratio test is also discussed.

cancelled ~~Wam1PIIT3~~

Estimating the Largest Parameter from Uniform Distribution in the Presence of Outliers from Generalized Uniform Distribution

Sushmita Jain, Ulhas J Dixit, Alladi Subramanyam

Indian Institue of Technology, India

Let X1,...,Xn be n observations, where k of these are outliers coming from generalized uniform distribution and the remaining n-k follow uniform distribution. Let θ_i be the unknown parameter associated with Xi. The problem of estimating the largest θ is considered. Let X(1)≥...≥X(n) denote the ordered observations. Suppose the population corresponding to X(1) is selected, and θ(i) denotes the parameter associated with X(i), 1≤i≤n. In this paper, we consider the estimation of θ(1) under the squared error loss $L(t,\theta) = (t-\theta)^2$. We construct estimators of θ(1) that dominate the natural estimators, by solving certain difference inequalities.

Wpm1PIIT4

Estimates and Confidence Bounds for the Number of False Hypotheses: A Partitioning Approach

Klaus Strassburger

Heinrich Heine University, Germany

The construction of estimates and confidence bounds for the number m_1 of false hypotheses in a given set of hypotheses $H_1,...,H_n$, is a challenging task with a wide field of applications ranging from meta-analysis and the development of powerful multiple test procedures to prevalence-estimates of a disease via diagnostic tests and quality measurement in genome-wide association studies.

In the first part of the talk, we present a general method for constructing confidence bounds for m_1 which is based on the partitioning principle. The basic idea of this approach is to test the disjoint partitioning hypotheses Ji, stating that the unknown parameter m_1 equals i, i=1,...,n, each at full level alpha. The index set of retained partitioning hypotheses then forms a (1-alpha) 100% confidence set for m_1. Thereby an index i that maximizes the p-value for testing J_i is a reasonable estimate for m_1.

In the second part of the talk several statistics for testing the partitioning hypotheses are proposed and the resulting estimates and bounds for m_1 are derived. Moreover, it will be shown how to incorporate prior knowledge of the distribution of the statistics (p-values) corresponding to false and true hypotheses H_i.

Wam1PIIIT1

A Sharp Upper Bound for the Expected Number of False Rejections

Alexander Gordon

University of North Carolina at Charlotte, USA

We present a method for calculating the exact level at which a multiple testing procedure controls the per family error rate (PFER), that is, the expected number of false rejections, under a general and unknown dependence between the p-values associated with the hypotheses being tested. We assume that the procedure is symmetric and satisfies the key assumption of monotonicity: reduction in some (or all) of the p-values can only increase the number of rejections. Our method applies, in particular, to the traditional stepwise procedures and leads to explicit formulas expressing the exact level of control of the PFER in terms of the procedure's thresholds (critical values).

Wam1PIIIT2

Control of the Expected Number of False Rejections in Multiple Hypotheses Testing

Marsel Scheer, Helmut Finner

Heinrich Heine University, Germany

The expected number of false rejections (ENFR) can be viewed as an important characteristic of any multiple testing procedure. However, compared to other error rate criteria like the popular familywise error rate (FWER) and the meanwhile even more popular false discovery rate (FDR) proposed in Benjamini and Hochberg (1995), control of ENFR seems to play no crucial role in multiple testing although Spjotvoll (1972) developed some optimality theory with respect to ENFR nearly forty years ago. The reason may be that ENFR control as investigated in Spjotvoll (1972) leads to Bonferroni type procedures which have the flavor of being too conservative. Some interesting results on a related error measure, that is the expected (type I) error rate defined by EER=ENFR/n with n denoting the number of all null hypotheses, can be found in Finner and Roters (2001, 2002, 2007).

In this talk we investigate a more flexible ENFR criterion and its potential for error rate control. We first show that there is a strong link between ENFR control and FDR if all p-values under null hypotheses are independently distributed. One of the main results is that in the class of step-up procedures based on a certain variant of the asymptotically optimal rejection curve (AORC) introduced in Finner et al. (2009) there exists a fundamental equivalence between FDR control and ENFR control. Moreover, it is shown that a step-down procedure based on the AORC, which is known to control the FDR, controls the ENFR, too. Further well-known linear stepwise procedures are investigated with respect to its ENFR behavior. Thereby, we provide formulas and asymptotic properties of the distribution of the number of false rejections. Then we investigate why FDR control by itself may lead to an inflated ENFR especially under dependence, while ENFR control by itself may lead to an inflated FDR. A reasonable compromise may be to control FDR and ENFR simultaneously with respect to suitable bounding functions.

Benjamini, Y. and Hochberg, Y. (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Stat. Methodol. 57, 289-300.

Finner, H., Dickhaus, T. and Roters, M. (2007). Dependency and false discovery rate: Asymptotics. Ann. Stat. 35, 1432-1455.

Finner, H., Dickhaus, T. and Roters, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. Ann. Stat. 37, 596-618.

Finner, H. and Roters, M. (2001). On the false discovery rate and expected type I errors. Biom. J. 43, 985-1005.

Finner, H. and Roters, M. (2002). Multiple hypotheses testing and expected type I errors. Ann. Stat. 30, 220-238.

Finner, H. and Scheer, M. (2011). Control of the expected number of false rejections in multiple hypotheses testing. Submitted for publication.

Spjotvoll, E. (1972). On the optimality of some multiple comparison procedures. Ann. Math. Statist. 43, 398-411.

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, 187-205.

Wam1PIIIT3

A Sufficient and Necessary Condition on Strong Control Generalized Familywise Error Rate in Multiple Hypothesis Testing

Huajiang Li

Allergan Inc., USA

In this talk we will present a sufficient and necessary condition on strong control generalized familywise error rate in multiple hypothesis testing. We will apply the condition to construct in some sense optimal level alpha multiple tests. A generic algorithm on searching optimal multiple tests will be provided.

Wam1PIIIT4

Simultaneous Confidence Regions for Closed Tests, including Hochberg’s and Hommel’s Procedures based on P-values

Olivier Guilbaud

AstraZeneca R&D, Sweden

Simultaneous confidence regions have previously been derived for various closed-testing procedures (CTPs), including a large class of Bonferroni-based CTPs, and a very useful subclass of graphical procedures where parts of alpha are successively recycled after rejections. This subclass includes many common procedures, e.g. Holm’s step-down procedure, fallback procedures, and gatekeeping procedures. These developments were recently extended (Guilbaud & Karlsson, 2011) to certain CTPs that are not Bonferroni based and which may utilize dependencies not utilized by Bonferroni-based procedures.

However, the derivation of confidence regions for other multiple testing procedures of practical interest has remained an unsolved problem. This is the case e.g. for Hochberg’s step-up procedure and Hommel’s more powerful procedure that is neither a step-up nor a step-down procedure. I will briefly discuss these previous developments, and then show how they can be further extended to cover more general CTPs, including a certain class of CTPs based on ordered p-values. Hochberg’s and Hommel’s procedures belong to this class. Interestingly, the confidence regions derived for these two procedures are closely related to those derived previously for Holm’s step-down procedure.

Guilbaud, O. and Karlsson, P. (2011). Confidence regions for Bonferroni-based closed tests extended to more general closed tests. Journal of Biopharmaceutical Statistics 21, 682-707.

Wam2PIIT1

Calculation of Simultaneous Confidence Intervals by Constraint Propagation

Georg Gutjahr, Frank Bretz

University of Bremen, Germany

In this talk, we describe an efficient algorithm to calculate confidence intervals of minimal length that are consistent with a given single-step test procedure. We assume that the unadjusted p-value functions of the individual tests can be extended to inclusion functions in interval-arithmetic and we briefly review the methods that are available to obtain such inclusion functions. We then express the confidence intervals as solution of a system of interval constraints and we use constraint propagation to solve this system and thereby obtain the simultaneous confidence intervals. As example, we discuss the calculation of simultaneous confidence intervals for ratios of means of normal distributions.

Wam2PIIT2

Multiple Comparisons among Components of Mean Vector under an Elliptical Population

Sho Takahashi, Takahiro Nishiyama, Takashi Seo

Tokyo University of Science, Japan

We consider multiple comparison procedure among components of mean vector under an elliptical population. In this talk, we discuss the simultaneous confidence intervals for multiple comparisons among mean components. In order to construct the simultaneous confidence intervals, it is required to derive the upper percentiles of maximum type statistic. However, it is difficult to find the exact values of the upper percentiles even under multivariate normality. So, approximations of the statistic based on Bonferroni's inequality are given by asymptotic expansion procedure. Furthermore, we investigate the effects of nonnormality on the upper percentiles of the statistic in elliptical distributions.

Wam2PIIT3

Conservative Simultaneous Confidence Intervals for Multiple Comparisons of Correlated Mean Vectors with a Control

Takahiro Nishiyama

Tokyo University of Science, Japan

We consider the simultaneous confidence intervals for multiple comparisons with a control among mean vectors from the multivariate normal distributions.

We discuss the approximate simultaneous confidence procedure proposed by Seo (1995) which concerning to the multivariate Tukey-Kramer procedure.

Seo (1995) conjectured that this procedure always construct the conservative approximate simultaneous confidence intervals.
We give the affirmative proof of this conjecture and give the upper bound for the conservativeness of this procedure in the case of five correlated mean vectors.

Finally, numerical results by Monte Carlo simulation are given.

Wam2PIIT4

On the Null-Problem in Multiple Hypotheses Testing

Veronika Gontscharuk, Helmut Finner

Heinrich Heine University, Germany

Suppose we are concerned with a multiple test problem with some dependence structure between test statistics or p-values. In general, depending on the underlying error rate criterion, dependency may increase or decrease the chance of false rejections.

We discuss this issue and give various illustrative examples. A special type of dependence is weak dependence. In terms of p-values, weak dependence appears if the empirical cumulative distribution function of p-values under null hypotheses is asymptotically stochastically bounded by the cdf of a uniform variate on [0,1]. There is some evidence that a multiple test procedure with asymptotic control of the false discovery rate (FDR) under independence also controls the FDR under weak dependence, provided the asymptotic threshold is bounded away from 0. But if the asymptotic threshold tends to 0, the situation is completely unclear. The question arises whether the FDR is controlled in such cases. We call this issue, which is often ignored in the literature, the null-problem in multiple hypotheses testing. It will be shown that weak dependence provides no guarantee of asymptotic FDR control if the null-problem appears.

Wam2PIIIT1

Adjusting for Multiple Testing Dependence via RIPOD

Sun Yunting, Nancy Zhang, Art B.Owen

Stanford University, USA

Most statistical methods for performing multiple testing rely on independence or some form of weak dependence among the data corresponding to the variables being tested. However, high dimensional studies rarely involve the analysis of independent variables with independent samples because of the presence of latent factors that comes from batch effect and population stratification. A latent factor not orthogonal to the primary predictors can lead to spurious association. We propose a method called RIPOD to tackle this issue by exploiting the sparsity of signals and low dimensionality of latent factors. Simulation studies show that our method has better power than existing methods such as SVA and EIGENSTRAT under most circumstances. Applying our method on Agemap mice gene expression data reveals some interesting relationship backed by and also contributing to the existing literature.

cancelled ~~Wam2PIIIT2~~

Hierarchical Testing of Subsets of Hypotheses

Marina Bogomolov, Yoav Benjamini

Tel Aviv University, Israel

As the size of large testing problems encountered in practice keeps increasing, more of these problems have further structure where the set of hypotheses can be partitioned into subsets of the hypotheses, and a discovery of some signal in a subset is of interest on top of the discovery of a signal in each of the many hypotheses on its own. Furthermore, the true state of the tested signals tends to be more similar within these subsets than across the subsets. Examples are regions in the brain in functional MRI research, sets of genes in genomic research, or geographical areas in disease outbreaks monitoring. The challenges in the analysis of such multiple testing problems will be discussed, and previous efforts to address them will be reviewed. We then present a few new methods to control various aspects of the False Discovery Rate, and discuss their benefits and limitations.

Wam2PIIIT3

Conservative Adjustment of *q*-value

Yinglei Lai

The George Washington University, USA

q-value is a widely used estimate of false discovery rate (FDR), which is a significance measure in the statistical analysis of recent high-throughput data. Unlike the traditional *p*-value, *q*-value is a random variable and may have a considerable variance, particularly when the permutation procedure has to be used for *p*-value calculations. An underestimated FDR can lead to unexpected false discoveries in a follow-up experimental validation. This issue has not been well addressed in the statistical literature. In this study, we suggest a conservative adjustment approach and we give a simple solution to calculate a conservative upper confidence limit of *q*-value.

Wam2PIIIT4

Communicating Advanced Multiple Test Strategies to Clinical Teams - a Case Study

Frank Bretz, Willi Maurer, Ekkehard Glimm

Novartis, Switzerland

Methods for addressing multiplicity issues have attracted much attention in the statistical literature over the past twenty years. Recent developments in this area include new classes of multiple test procedures, such as fixed-sequence, fallback and gatekeeping procedures. Graphical approaches have been introduced to visualize and communicate some of these advanced multiple test strategies. In this presentation we describe in the context of a real case study how clinical teams can be engaged at the planning stage of a trial to elicit complex importance relationships between competing study objectives and how these can be reflected when constructing a suitable multiple test strategy.

Wpm1PIT1

A Multiple Comparison Procedure for Hypotheses with Gatekeeping Structure

Xiaolong Luo, S. Peter Ouyang

Celgene Corporation, USA

In this paper, we develop a multiple comparison procedure for hypotheses with gatekeeping structures. This procedure will utilize the correlation among individual test statistics without making any parametric assumption. We derive a general asymptotic multivariate normal distribution of all involved test statistics, obtain an estimation of the correlation matrix and utilize a recently developed computing procedure "pmvnorm" to calculate the operating characteristics. We construct the closed test procedure based on the gatekeeping relationship and the asymptotic multivariate normal distribution. Simulation analyses will be used to illustrate its relative advantage compared with popular Bonferroni and truncated Hommel procedures. A trial example is used to illustrate its application.

Wpm1PIT2

On Validity of Analysis of Mortality

Qing Liu

Janssen Pharmaceutical Companies of J&J, USA

For drug development of life-threatening disease, mortality is often not used as a primary endpoint because the sample size required to detect a meaningful treatment difference is usually prohibitively large. Instead, a clinical endpoint measuring a more immediate treatment effect on certain important aspect of disease is used as a primary endpoint. Mortality is either listed as a secondary or exploratory endpoint. There are examples, however, that a new drug is shown to reduce mortality rate and yet fails to demonstrate a statistically significant effect on a primary endpoint. Following the conventional hierarchical multiple testing procedure, it is commonly believed that any statistical significance of the mortality endpoint cannot be declared without inflating the multiple type 1 error rates. We point out that such a multiple testing procedure does not taken into account implicit evaluations of totality of evidence which without exception are integral part of regulatory review and decision process. We unveil such an implicit regulatory framework and show in fact that mortality endpoint can meaningfully analyzed and interpreted even after analysis of primary endpoint fails to demonstrate statistical significance. This exonerates the U. S. Food and Drug Administration (FDA) of its past regulatory precedence. We also show how this regulatory framework can be applied to clinical trials employing group sequential or adaptive designs.

Wpm1PIT3

Statistical Consideration for the Design and Analysis of Clinical Trials with Targeted Subgroups

Mohamed Alosh, Mohammad Huque

U.S. Food and Drug Administration, USA

It is common practice to examine efficacy results by subgroups with the objective of learning of differential treatment effects among subgroups. Such analyses are frequently carried out although it is well known that there are several shortcomings of the findings of such post-hoc analyses, including the possibility that they might be driven by chance alone, thus limiting their utility. However, recognizing that the treatment effect might vary by subgroups, which can be characterized by biological or genomic factors, there has been growing interest in designing clinical trials with the objective of establishing an efficacy claim for the total population and/or targeted subgroups. For this objective, a proper study design and analysis plan needs to be put in place a priori to ensure control of the Type I error rate and meaningful interpretation of study findings, so that efficacy findings for targeted subgroups are trustworthy. In this presentation we discuss several statistical concepts related to the design and analyses of such trials including: (i) probability of positive, as well as negative, chance findings for subgroups, (ii) power interplay between subgroups and total population along with consideration of enrichment designs, (iii) multiple testing strategies for targeted subgroups and total population, which ensure meaningful interpretation of study findings. Finally, we consider application of these concepts to the clinical trial setting.

Wpm1PIT4

Multiple Comparisons in the ANOVA Model under Heteroscedasticity

Gerhard Hommel

Universitätsmedizin Mainz, Germany

It is well known that the classical t test can become very conservative as well as anticonservative for an unbalanced design when the variances are unequal. It is therefore recommended to use the approximative solution by Welch (1938).

When more than two samples are compared within the ANOVA model, less investigations have been made, but one can expect a similar behavior of global and/or multiple tests in the case of heteroscedasticity. The results of a simulation study are described where different types of the closure test were used. In particular it was investigated how common variance estimates (pooled over all samples or separately for pairs of samples) influence the type I error rates.

When heterogeneous variances are expected, it is recommended to avoid the use of pooled variance estimates, even in the balanced case. Instead, one should use Welch t tests (or Welch F tests) within the closure test.

Wpm1PIIT1

Heteroscedastic Analysis of Means with Unequal Sample Sizes

Miin-Jye Wen

National Cheng Kung University, Taiwan

The heteroscedastic analysis of means (HANOM) is a testing procedure for comparing a group of means to see if any of them are significantly different from the overall mean with unequal variances. In practice, equal sample sizes often are not practical due to time saving or budget limitation, etc. Hence method for dealing with unequal sample sizes is necessary. When the population variances are unknown and unequal, Dudewicz and Nelson (2003) proposed a design-oriented two-stage procedure for HANOM, which requires additional samples at the second stage. If obtaining more samples is not practical or feasible, we use a data-oriented single-stage sampling procedure, which originally developed by Chen and Lam (1989), to test the null hypothesis in HANOM models with unequal sample sizes. It does not require extra samples, and can reach a conclusion much easier and save time and budget. A example is given to illustrate how this procedure works.

Wpm1PIIT2

Multiple Testing of Composite Null Hypotheses in Heteroscedastic Models

Alexander McLain, Wenguang Sun

National Institutes of Health, USA

In large-scale studies, the true effect sizes often range continuously from zero to small to large, and are observed with heteroscedastic errors. In practical situations where the failure to reject small deviations from the null is inconsequential, specifying an indifference region can greatly reduce the number of unimportant discoveries in multiple testing. In addition, it is desirable to address the heteroscedasticity issue for valid and efficient simultaneous inference. We show that the composite null hypotheses and heteroscedasticity of errors lead to new concepts of the null distribution in large-scale multiple testing. We propose a mixed deconvoluting kernel method for estimating the null distribution that include both zero and small effects, and then develop an optimal procedure for testing composite nulls that minimizes the false non-discovery rate subject to a constraint on the false discovery rate. The proposed approach is different from conventional methods in that the effect size, statistical significance and multiplicity issues are addressed integrally. The new features and advantages of our approach are demonstrated using both simulated and real data. The numerical results show that our new procedure enjoys superior performance by effectively eliminating nonessential discoveries and yielding more reproducible scientific results.

Wpm1PIIT3

MCP under Stochastic Order: Controlling FDR

Jinde Wang, Jianwei Gou

Nanjing University, China

Multiple comparison of the effects of several treatments with the control(MCC for short) has been a central problem in many areas and it has been studied quite a lot. However nearly all of existing papers are devoted to comparing the means (or other location parameters) of these effects. Means usually provide only a part of the whole information of the random variables(here,effects). By comparing the distributions of random variables it can be expected to get more useful and deeper conclusion . It is a MCP under stochastic order. Multiple comparison under stochastic orders has almost not been studied.

This paper gives a way to compare the distributions for MCC problems, controlling the false discovery rate. The test controlling FDR under stochastic order faces more challenges than the tests controlling FDR in most of traditional cases. A hard one is that the *p*-values required for the test can't be assumed independent,as was done in many papers. Moreover,in order to make our results applicable more widely, we do not assume the random variables involved here are normal. This makes distributions of the *p*-values more difficult to be found. We report how to solve these problems.

Wam2PIIIT2

Controlling the False Discovery Rate in Two-Stage Combination Tests for Multiple Endpoints

Sanat Sarkar

Temple University, USA

We consider the problem of testing hypotheses associated with multiple endpoints in a two-stage adaptive design setting in which the hypotheses are screened at the first stage based on some rejection and acceptance thresholds for the p-values and the follow-up hypotheses are tested having combined the p-values from the two stages. For this problem, we will present two BH type methods to control the false discovery rate (FDR), extending the original BH method and its adaptive version from single-stage to a two-stage setting. These methods will be shown to theoretically control the FDR under independence. Considering two types of combination function - Fisher's and Simes' - and three types of dependence - equal, clumpy and AR(1) - for the underlying test statistics, we will provide numerical evidence that these methods can potentially control the FDR even under certain dependence situations. Application of these methods to a real data set will also be presented.

Wpm1PIIIT1

Adaptive FWER and FDR Control under Block Dependence

Wenge Guo, Sanat, Sarkar

New Jersey Institute of Technology, USA

Recently, a number of adaptive multiple testing procedures have been proposed. However, in a non-asymptotic setting, the FWER or FDR control of these adaptive procedures is only proved under independence, although simulation studies have suggested that these procedures perform well under certain dependence. In this talk, several variants of the conventional adaptive Bonferroni and Benjamini-Hochberg methods will be presented, along with proofs that these procedures provide ultimate control of the FWER or FDR under block dependence. Results of simulation studies comparing the performances of these adaptive procedures with the conventional FWER and FDR controlling procedures under different types of dependence will also be presented.

Wpm1PIIIT2

Exact Calculations for the False Discovery Proportion and Applications

Etienne Roquain, Gilles Blanchard, Thorsten Dickhaus, Fanny Villers

Universite Pierre et Marie Curie, France

We provide exact formulas for the distribution of the false discovery proportion (FDP) and for the false discovery rate (FDR) of step-up, step-down and step-up-down procedures, with an arbitrary rejection curve. The *p*-values are assumed either independent or coming from an equicorrelated multivariate normal model (with a common alternative). The set of true null hypotheses is either fixed or comes from a mixture model.

In any case, our formulas can be fully computed numerically using Steck's type recursions (for reasonably large m, say m<=1000 under the mixture or <=100 without mixture). This new approach is useful to:

* investigate new theoretical results which are of interest in their own right, related to least/most favorable configurations for the FDR;

* avoid cumbersome simulations, for instance, to check numerically if a given procedure controls the FDP/FDR.

(This is a joint work with Gilles Blanchard, Thorsten Dickhaus and Fanny Villers, which has been partly published under the reference ``E.Roquain and F.Villers. Exact calculations for false discovery proportion with application to least favorable configurations. Ann. Statist. 39(1):584--612, 2011.")

Wpm1PIIIT3

Generalized Stepwise Procedures to Control the False Discovery Rate

Scott Roths

Pennsylvania State University, USA

Among the most popular procedures used to control the false discovery rate (FDR) in large-scale multiple testing are the stepwise ones, where the marginal p-values are ordered and compared with specified cutoffs, according to a stopping rule. Starting with the most significant p-value, the stepdown procedure rejects each null hypothesis as long as the corresponding cutoff is not exceeded. The stepup procedure starts with the least significant p-value and proceeds in the opposite direction and accepts each null hypothesis, provided the p-value does not exceed its cutoff. These procedures have been shown to control the FDR under certain types of dependancies, including the kind relevant to multiple comparisons with a control. This talk discusses a proposed method that generalizes these stepwise procedures by allowing the stepdown procedure to continue rejecting as long as the fraction of p-values not exceeding their cutoffs is sufficiently large. The stepup procedure is similarly generalized. For appropriate choices of this fraction bound, increased power may be obtained without compromising FDR control, particularly when the non-null p-values are more extreme. The proposed method is illustrated with simulated and real data.

Wpm1PIIIT4

Designing Multi-Regional Clinical Trial with Different Regional Required Primary Endpoints

Yi Tsong, Chin-Fu Hsiao, Hsiao-Hui Tsou, Wan-Jung Chang, Xiaoyu Dong

U.S. Food and Drug Administration, USA

One of the challenges of multi-regional drug development program is to design and analyze a multiple regional clinical trial with the objective to satisfy different regional requirements on primary endpoint. Considering a multi-regional clinical trial (MRCT) designed to test for two different primary endpoints required by two different regions, data of a regular well controlled parallel arm trial will be used to test for two null hypotheses in terms of two distinct yet highly correlated endpoints. The two hypotheses may be tested sequentially or simultaneously. Depending on the structure of the hypotheses to be tested and the understanding of type I error rate controlling, various scenarios of type I error rate adjustments may be applied. Further more, for the objectives of getting approval of regional authorities of different primary endpoints, various sample size and power adjustment for multiple comparisons may be applied. In this presentation, comparisons of different approaches are discussed systematically.

Tham1PIT1

Test Procedures for the Assessment of the Components of Composite Endpoints

Geraldine Rauch, Meinhard Kieser

University of Heidelberg, Germany

Composite endpoints are increasingly used in clinical trials, particularly in the field of cardiology. Thereby, the overall impact of the therapeutic intervention is captured by including several events of interest in a single variable. In the ICH E9 Guideline, it is stated that 'this approach addresses the multiplicity problem without requiring adjustment to the type I error' (ICH, 1999). In fact, to demonstrate the significance of an overall clinical benefit, it is sufficient to assess the test problem formulated for the composite. However, even if a statistically significant and clinically relevant superiority is shown for the composite endpoint, there is the need to evaluate the treatment effects for the constituting components. For example, the Points to Consider on Multiplicity (2002) require that "… if all cause mortality is a component, a separate analysis of all cause mortality should be provided to ensure that there is no adverse effect on this endpoint."

We propose multiple test procedures that enable decisions about the components of a composite endpoint under control of the type I error rate. The properties of these approaches in terms of required sample size and power are compared. It is shown how the issue of follow-up decisions about the components can be addressed in the planning stage. Application is illustrated by a clinical trial example.

Tham1PIT2

Statistical and Regulatory Consideration for Multi-Item Patient Reported Outcome (PRO)

Rima Izem, Mohammad Huque

U.S. Food and Drug Administration, USA

This presentation will discuss some statistical considerations for development and use of multi-item PROs in clinical trials.

As for many composite endpoints, multiple testing issues arise in clinical studies with multi-item PROs. Multi-item PROs are composite endpoints since answers to several items or questions are combined to derive a PRO score. Although items may be measuring different signs and symptoms, a combined PRO score is meaningful when all items measure one construct [as outlined in PRO guidance issued in December 2009]. Multiple testing issues arise when not only the total PRO score is of interest but also the scores for individual items or subgroups of items.

Our talk will show some examples of how multi-item PROs are used in clinical trials and how their results are displayed in labels. We will also discuss proposals to control for multiple testing for this special case of composite endpoints.

Tham1PIT3

Establishing Non-Inferiority and Equivalence in Matched Pair Designs with Multiple Endpoints based on McNemar's Test

Jin Xu, Menggang Yu

East China Normal University, China

We propose a general method for testing non-inferiority or equivalence in matched pair designs with multiple endpoints. The method employs intersection-union principle on the marginal score statistics to obtain an asymptotic α-level test. Power and sample size calculation are obtained by a simple numerical method that takes into account the correlation structure among the endpoints.

A two-stage adaptive design with internal pilot study is also proposed for sample size re-estimation when the nuisance parameters are not available. Both blinded and unblinded re-estimation methods are considered. The proposed methods are evaluated by simulation studies and applied to a cancer quality of life trial.

Tham1PIT4

Development of Gatekeeping Procedures in Confirmatory Trials

Alex Dmitrienko

Eli Lilly and Company, UK

This presentation introduces general guidelines for the development of multiple testing procedures (known as gatekeeping procedures) in confirmatory clinical trials with multiple families of objectives, e.g., trials with primary and secondary endpoints, dose-placebo comparisons and patient populations. Gatekeeping procedures are used in hypothesis testing problems of this kind to control the Type I error rate across the multiple families and have attracted much attention in clinical drug development. The general process for building gatekeeping procedures includes identification of candidate procedures that are consistent with logical relationships among the multiple objectives, utilization of all available distributional information and selection of most powerful procedures based on application-specific criteria. The principles discussed in this presentation are illustrated using a clinical trial in patients with Type II diabetes.

Tham1PIIT1

Multistage Parallel Gatekeeping with Retesting

George Kordzakhia, Alex Dmitrienko

U.S. Food and Drug Administration, USA

This talk introduces a general method for constructing multistage parallel gatekeeping procedures with a retesting option. This approach serves as an extension of general multistage parallel gatekeeping procedures (Dmitrienko, Tamhane and Wiens, 2008) and parallel gatekeeping procedures with retesting for two-family problems (Dmitrienko, and Tamhane, 2011). It was shown in the latter paper that power of parallel gatekeeping procedures can be improved by adding an additional retesting stage which enables retesting of the primary null hypotheses using a more powerful component procedure than the original procedure if all secondary null hypotheses are rejected. The new method enables clinical trial researchers to construct a class of more general multistage parallel gatekeeping procedures with retesting. The new procedures support multiple retesting of the primary and secondary families and do not require all secondary null hypotheses be rejected. (This talk will be given in the session entitled "Analysis of clinical trials with multiple objectives" organized by Brian Wiens.)

Tham1PIIT2

Reproducibility of Conclusions on Multiple Hypotheses from One or More Families

Brian Wiens, Alex Dmitrienko

Alcon Laboratories, Inc., USA

We consider analysis of two identical pivotal trials with correlated multiple endpoints evaluated by the fixed sequence, weighted Holm or fallback procedure for a single family of hypotheses, or by a gatekeeping strategy for multiple families of hypotheses. For approval, at least one endpoint must be significant in both studies. Evaluation of the procedures as closed tests distinguishes which has better power in different situations, and simulations distinguish aspects of aesthetics such as the probability of obtaining inconsistent results in the two studies. Testing strategies that allow for flexibility in determining the hypotheses that are rejected often provide the most appealing strategies, both for power and for aesthetics. We propose that such evaluations should be a routine part of the process for planning a phase III confirmatory strategy.

Tham1PIIT3

Sequentially Rejective Graphical Multiple Test Procedures with Memory

Willi Maurer, Frank Bretz

Novartis Pharma AG, Switzerland

The graphical sequentially rejective test procedure by Bretz et al. (2009) propagates the local significance level of a rejected hypothesis to the not yet rejected hypotheses according to pre-specified transition weights. In the graph defining the test procedure, hypotheses and their local significance levels are represented by weighted vertices and the transition weights by weighted directed edges. An algorithm provides the rules for updating the significance levels of the vertices and the transition weights of the edges after a rejecting an individual hypothesis. This graphical procedure has no memory in the sense that the origin of the propagated significance levels is ignored in subsequent iterations. Memory would allow the further propagation of significance levels to be dependent of their origin and thus reflect the grouped parent-descendant structures of the hypotheses. In some clinical trial applications, memory is desirable to address properly such a dependence structure in families of hypotheses. We will give examples of such situations and show how memory can be induced by convex combination of graphical procedures. The resulting entangled graphs provides an intuitive way to represent the underlying partial order and relative dependence of the hypotheses to be tested, is as easy to perform as those based on single graphs, is sequentially rejective and allows strong control of the Type I error rate.

Bretz F, Maurer W, Brannath W, Posch M. (2009) A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine 8, 586-604.

Tham1PIIT4

Cherry-Picking? Multiple Testing for Exploratory Research

Jelle Goeman, Aldo Solari

Leiden University Medical Center, The Netherlands

Motivated by the practice of exploratory research, we formulate an approach to multiple testing that reverses the conventional roles of the user and the multiple testing procedure. Traditionally, the user chooses the error criterion, and the procedure the resulting rejected set. Instead, we propose to let the user choose the rejected set freely, and to let the multiple testing procedure return a confidence statement on the number of false rejections incurred. In our approach, such confidence statements are simultaneous for all choices of the rejected set, so that post hoc selection of the rejected set does not compromise their validity. The proposed reversal of roles requires nothing more than a review of the familiar closed testing procedure, but with a focus on the non-consonant rejections that this procedure makes. We suggest several shortcuts to avoid the computational problems associated with closed testing.

Tham1PIIIT1

Partitioning Testing for Broad Efficacy and Efficacy in Genomic Subgroups

Szu-Yu Tang, Yi Liu, Jason C. Hsu

The Ohio State University, USA

Traditionally, clinical studies aim to prove drug efficacy over broad patient populations. However, in addition to testing for broad efficacy, there may be reason to test for efficacy in one or more genomic subgroups.

For example, there is biological reason and preliminary data suggesting breast cancer patients with BRCA mutations might benefit from PARP drugs more. As another example, mutations in P450 metabolizer genes such as 2C19 and 2D6 might affect efficacy of some drugs.

We first show partition testing simplifies the formulation of multiple testing for such studies. We then show partition testing is powerful in executing the analysis, by first comparing weak partitioning test against Hochberg's step-up test, and then make further improvement by applying the strong partitioning principle.

Tham1PIIIT2

Correction of the Significance Level after Multiple Coding of an Explanatory Variable in Generalized Linear Model.

Jérémie Riou, Benoit Liquet

Inserum U897- Isped, France

In statistical modelling, finding an optimal encoding for an exploratory quantitative variable implies many tests. This process involving multiple testing problems requires the correction of significance level.
First, we focus on the context of a binary coding for a generalized linear model. For each coding, a test on the nullity of the coefficient associated with the new coded variable is computed. The selected coding corresponds to the one associated with the largest statistical test. In this context, it exists an exact correction of test significance level [1] associated with the maximum of the tests. This correction is compared to Bonferroni and Efron corrections and also to resampling procedures such as permutation and bootstrap [2].

Second, we focus on categorizing the continuous variable in more than two classes (m classes). For each coding, a test on the nullity of coefficient vector associated with the categorical variable is computed. In this context, the statistical score test follows asymptotically a chi-square distribution with m-1 degrees of freedom. To accurately calculate the significance level of the maximum of the tests, it is necessary to know the distribution function of a multivariate chi-square. Several methods have been developed to approximate this distribution, however they require assumptions that do not fit our context. Therefore, we propose to determine the significance level of the maximum of the tests by resampling methods such as permutation, parametric bootstrap and the Stochastic Approximation in Monte-Carlo computation (SAMC) algorithm. The SAMC algorithm has recently been developed for avoiding computational time. This algorithm is also relevant for multiple testing [3], [4].

Finally, these methods are compared in a simulation study and further applied on real data. An R package for their implementation is also proposed.

Keywords: Adjusted p-value, Bonferroni, Bootstrap, Generalized Linear Model, Multiple coding, Permutation, Resampling procedure, SAMC, Significance level.

[1] B.Liquet and D.Commenges (2004). Computation of the p-value of the

minimum of score tests in the generalized linear model, application to multiple coding. Statistics & Probability Letters, 71:33-38.
[2] PH.Westfall and SS.Young (1993). Resampling-based Multiple Testing. Wiley.
[3] K.Yu, F.Liang, and J.Ciampa (2011). Efficient p-value evaluation for resampling based tests. Biostatistics, 0:1-11.
[4] F.Liang, C.Liu, and RJ.Caroll (2007). Stochastic approximation in monte carlo computation. Journal of american Statistical Association, 102:305-320.

Tham1PIIIT3

Stability Based Testing for the Analysis of fMRI Data

Joke Durnez, Beatrijs Moerkerke

Ghent University, Belgium

Neurological imaging has become increasingly important in the field of psychological research. The leading technique is functional magnetic resonance imaging (fMRI), in which a correlate of the oxygenlevel in the blood is measured (the BOLD-signal). In an fMRI-experiment, a time series of brain images is taken while participants perform a certain task. By comparing different conditions, the task-related areas in the brain can be localised. An fMRI study leads to enormous amounts of data. To analyse the data adequately, the brain images are devided into a large number of volume units (or voxels). Subsequently, a time series of the measured signal is modelled voxelwise as a linear combination of different signal components, after which an indication of activation can be tested in each voxel. This encompasses an enormous number of simultaneous statistical tests (+/-250 000 voxels). As a result, the multiple testing problem is a serious challenge for the analysis of fMRI data.

In this context, classical multiple testing procedures such as Bonferroni and Benjamini-Hochberg (Benjamini & Hochberg, 1995) have been applied to respectively control the family-wise error rate (FWER) and the false discovery rate (FDR)(Genovese, Lazar, & Nichols, 2002). Random Field Theory (Worsley, Evans, Marrett, & Neelin, 1992) controls the FWER while accounting for the spatial character of the data. Because of the dramatically decrease in power when controlling the FWER, methods to control the topological false discovery rate (FDR) were developed (Chumbley & Friston, 2009; Heller, Stanley, Yekutieli, Rubin, & Benjamini, 2006).

A general shortcoming of current procedures is the focus on detecting non-null activation while a non-null effect is not necessarily biologically relevant. Moreover, failing to reject the hypothesis of no activation is not the same as confidently excluding important effects. Another aspect that remains largely unexplored is the stability of test results which can be defined as selection variability of individual voxels (Qiu, Xiao, Gordon, & Yakovlev, 2006).

Given the need to control both false positives (type I errors) and false negatives (type II errors) in a direct manner (Lieberman & Cunningham, 2009), we approach the multiple testing problem from a different angle. Following the procedure of (Gordon, Chen, Glazko, & Yakovlev, 2009) in the context of gene selection, we present a statistical method to detect brain activation that not only includes information on false positives, but also on power and stability. The method uses bootstrap resampling to extract information on stability and uses this information to detect the most reliable voxels in relation to the experiment. The findings indicate that the method can improve stability of procedures and allows a direct trade-off between type I and type II errors. In this particular setting, it is shown how the proposed method enables researchers to adapt classical procedures while improving their stability. The method is evaluated and illustrated using simulation studies and a real data example.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57 (1), 289-300.

Chumbley, J., & Friston, K. (2009). False discovery rate revisited: FDR and topological inference using gaussian random &#64257;elds. NeuroImage, 44, 62-70.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage, 15, 870-878.

Gordon, A., Chen, L., Glazko, G., & Yakovlev, A. (2009). Balancing type one and two errors in multiple testing for differential expression of genes. Computational Statistics and Data Analysis, 53, 1622-
1629.

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., & Benjamini, Y. (2006). NeuroImage, 33, 599-608.

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and type II error concerns in fMRI research: rebalancing the scale. Social cognitive and affective neuroscience, 4, 423-428.

Qiu, X., Xiao, Y., Gordon, A., & Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis. bioinformatics, 7 (50).

Worsley, K., Evans, A., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for cbf activation studies in human brain. Journal of Cerebral Blood Flow Metabolism, 12(6), 900-918.

Tham1PIIIT4

Testing Multiple Endpoints in Complex Clinical Trial Designs

H.M. James Hung, Sue-Jane Wang

U.S. Food and Drug Administration, USA

In many disease areas, designs of pivotal clinical trials are increasingly complex. For assessing cardiovascular risks in a clinical program, multiple trials may be jointly analyzed to assess a mortality endpoint whereas each trial is planned to assess a different endpoint. For assessing a rare safety event, multiple trials may be jointly analyzed to assess collectively for sufficient study power and consistency across trials. In another case, a single trial may be conducted to assess a major adverse clinical event and a symptom endpoint by splitting the trial into two trials. Active controlled designs with or without a placebo arm and adaptive designs are also complex with many difficult problems for testing endpoints. This paper will present the challenges of the conventional statistical inference frameworks and stipulate a number of approaches to the multiplicity problems associated with testing multiple endpoints under such designs in confirmatory clinical trials.

Tham2PIT1

An Adaptive Extension of a Two-Stage Group Sequential Procedure for Testing a Primary and a Secondary Endpoint with Gatekeeping Constraint

Ajit Tamhane, Yi Wu, Cyrus Mehta

Northwestern University, USA

In this talk we present an adaptive extension of the two-stage group sequential procedure proposed in Tamhane, Mehta and Liu (2010) for testing a primary and a secondary endpoint where the primary endpoint serves as a gatekeeper for the secondary endpoint. That paper assumed a simple setup of a two-stage procedure and a bivariate normal distribution for the two endpoints with correlation coefficient ρ being either an unknown nuisance parameter or a known constant. Under the former assumption the least favorable value of $\rho=1$ was used which results in a conservative procedure. On the other hand, the latter assumption is unrealistic. A naive use of the sample correlation coefficient r from the first stage data in place of unknown ρ can lead to a liberal procedure. We show how an upper confidence limit on ρ can be used to accurately control the type I error rate without excessively sacrificing power that is implicit in the conservative procedure. Other adaptive extensions will be mentioned.

Tham2PIT2

A Nonparametric Procedure to Compare Clustered Multiple Endpoints

Aiyi Liu, Chunling Liu, Liansheng Tang

NIH, USA

In biomedical research such as epidemiology studies multiple outcomes are almost always measured and in many situations these outcomes form natural clustered. As an example, the healthy eating index (HEI), an important measure for management of diabetes, falls naturally into categories related to vegetable, fruit, meat, milk, etc. Where making comparisons in these clustered multiple outcomes using conventional test statistics, ignoring these clustering features may result in loss of power in testing hypothesis. In this talk I will present a nonparametric testing procedure that combines the rank-sum test statistics of O'Brien's (1984), and the max-statistics. Simulation studies show that the proposed procedure gains power in testing group difference in multiple endpoints when the outcomes within each cluster are directionally correlated. Healthy eating index data are used to evaluate the effect of family characteristics on the eating behavior.

Tham2PIT3

Graphical Approaches for Multiple Endpoint Problems using Weighted Parametric Tests

Ekkehard Glimm, Bretz Frank, Maurer Willi

Novartis Pharma AG, Switzerland

In clinical trials, the effect of a new treatment is often investigated in multiple endpoints, for example different patient characteristics (e.g. weight loss and change of HBA1C level), different doses of a drug, different time points at which the effect is measured, or combinations thereof. When the new treatment is compared to an established one or placebo, control of the familywise error rate is often required to avoid over-optimistic conclusions about the effect of the new treatment. In addition, usually some comparisons are more important than others, such that partial hierarchies of primary and secondary hypotheses arise. These challenges have triggered the development of stepwise multiple testing procedures, like the Bonferroni-Holm procedure and generalizations to gatekeeping and fallback procedures. Bretz et al. (2009, Statistics in Medicine 28, 586-604) have suggested a graphical approach that allows an easy, transparent description of such procedures by means of directed graphs. However, their paper restricts the investigation to Bonferroni-based procedures, i.e. methods that do not exploit knowledge about the multivariate distribution of corresponding test statistics.

The talk will first present the approach by Bretz et al. (2009) and then discuss its extension to situations where endpoints are (asymptotically) normally distributed with known correlations (which, for example, may occur when several doses of a new drug are compared with the same active control). The situation where all or some of the correlations have to be estimated from the data will also be considered.

Tham2PIT4

Resolving the Type I and Type II Error Dilemma for Clinical Safety Analyses

Devan Mehrotra, Adeniyi Adewale

Merck Research Laboratories, USA

Comparative analyses of safety/tolerability data from a typical phase III randomized clinical trial generate multiple p-values associated with adverse experiences (AEs) across several body systems. A common approach is to "flag" any AE with a p-value less than or equal to 0.05, ignoring the multiplicity problem. Despite the fact that this approach can result in excessive false discoveries (false positives), many researchers avoid a multiplicity adjustment in order to curtail the risk of missing true safety signals. We propose a new flagging mechanism that significantly lowers the false discovery rate (FDR) without materially compromising the power for detecting true signals, relative to the common no-adjustment approach. Our simple two-step procedure is an enhancement of the Mehrotra-Heyse-Tukey approach that leverages the natural grouping of AEs by body systems. We use simulations to show that, on the basis of FDR and power, our procedure is an attractive alternative to (i) the no-adjustment approach, (ii) a one-step FDR approach that ignores the grouping of AEs by body systems, and (iii) a recently proposed two-step FDR approach for much larger-scale settings like genome-wide association studies. An illustrative example is used to reinforce the key points.

Tham2PIIT1

Analysis of Multi-regional Clinical Trials: Applying a Two-Tier Procedure to Decision-Making by Individual Local Regulatory Authorities

Yunling Xu, Nelson, Lu

CDRH/US FDA, USA

The number of multi-regional clinical trials (MRCT) has been increasing for medical products development. However, the presence of inherent regional difference in treatment effect poses a great challenge to local regulatory decision-making. In face of this challenge, published literature so far has been focusing on assessment of treatment effect consistency across regions in MRCTs. Here, we propose a two-tier procedure for analyzing MRCT data for local regulatory decision making, allowing treatment effect varying from region to region. With the two-tier procedure, we differentiate direct evidence from extended evidence while using both to exemplify the advantage of MRCTs for decision making by local regulatory authorities. Use of the two-tier procedure is illustrated with examples of randomized controlled superiority trials of medical devices.

Tham2PIIT2

Multiple Testing with Latent Variable Model for Ordered Categorical Response

Tong-Yu Lu, Wai-Yin Poon, Siu Hung Cheung

College of Economics and Management, China Jiliang University, China

Ordered categorical data are frequently encountered in clinical studies. A popular method for comparing the efficacy of treatments is to use logistic regression with the proportional odds assumption. The test statistic is based on the Wilcoxon-Mann-Whitney test. However, the proportional odds assumption may not be appropriate. In such cases, the probability of rejecting the null hypothesis is much inflated even though the treatments have the same mean efficacy. We propose an alternative approach that does not rely on the proportional odds assumption when the responses can be conceptualized as manifestations of some underlying continuous variables. A latent normal distribution is utilized and under the latent variable model framework, we derive testing procedures that compare several treatments to a control. Both single-step and stepwise procedures are introduced and these procedures are compared based on their power.

Data from clinical trials are used to illustrate the proposed procedures.

Tham2PIIT3

Optimizing Drug Development; an Application to Diabetes

Zoran Antonijevic, Klas Bergenheim, Carl-Fredrik Burman, Martin Kimber, David Manner, Jose Pinheiro

Quintiles, USA

This presentation will discuss optimization of selected Phase II and Phase III design parameters and decision criteria such that regulatory and commercial outcomes are maximized. Impacts will be assessed at the program level, with primary outcome being the expected value of a product as measured by the expected NPV. Diabetes is the indication to which these methods will be applied. The existing regulatory guidance is strictly followed and realistic revenue models applied. Creation of a utility function for dose selection that is consistent with the expected revenues model will be described. Optimal sample sizes in Phase II and Phase III will also be discussed, as well as the application of an adaptive design.

Tham2PIIT4

Challenges in Developing Tailored Therapeutics to Improve Personalized Medicine

Steve Ruberg

Eli Lilly and Company, USA

With the advent of molecular biology and the genomic revolution, expectations are quite high for developiong medicines that are tailored to specific subgroups of patient in a disease population. Most often we seek to identify such subgroups by well defined, measurable characteristics of the patient. However, finding the right characteristics (i.e. biomarkers in a very broad sense) is now recognized as a very complicated endeavor fraught with many multiplicity issues. In order to market a drug to specific subgroups of patients defined by one or more biomarkers, the sponsor must provide evidence based on adequate and well-controlled trials in order to make appropriate drug labeling statements and advertising/promotional claims approved by the FDA or other regulatory bodies worldwide.

This talk will cover some of the difficulties faced by pharmaceutical companies developing tailored therapeutics in a regulated environment. While some practical issues as well as health care realities will be explored, the talk will focus on statistical issues with some review of some useful solutions and proposals for future implementation.

Tham2PIIIT1

A Novel Recursive Partitioning Method for Establishing Response to Treatment in Subpopulations

Ilya Lipkovich, Alex Dmitrienko, Jonathan Denne , Gregory Enas

Eli Lilly and Company, USA

We propose a novel recursive partitioning method for identifying subgroups of subjects with enhanced treatment effects based on a differential effect search algorithm. The idea is to build a collection of subgroups by recursively partitioning a database into two subgroups at each parent group, such that the treatment effect within one of the two subgroups is maximized compared to the other subgroup. The process of data splitting continues until a predefined stopping condition has been satisfied. The method is similar to “interaction tree” approaches that allow incorporation of a treatment-by-split interaction in the splitting criterion. However, unlike other tree-based methods, this method searches only within specific regions of the covariate space and generates multiple subgroups of potential interest. We develop this method and provide guidance on key topics of interest that include generating multiple promising subgroups using different splitting criteria, choosing optimal values of complexity parameters via cross-validation, and addressing Type I error rate inflation inherent in data mining applications using a resampling - based method. The operating characteristics of the procedure are evaluated using a simulation study and the method is illustrated with a clinical trial example

Tham2PIIIT2

Interaction Trees for Subgroup Analysis

Xiaogang Su, Xiaogang

University of Alabama at Birmingham, USA

The breast cancer education intervention (BCEI) study is a randomized controlled psychoeducational intervention trial aiming to improve the quality of life (QOL) of breast cancer survivors. Borrowing the idea of recursive partitioning and following the convention of classification and regression trees, an exploratory procedure, termed interaction trees, is proposed to understand better the differential effects of the BCEI on longitudinal quality-of-life data. The resultant tree model identifies several objectively defined subgroups: in some groups the BCEI is quite effective whereas in others it may not be. Based on the final tree structure, a permutation test is used to assess existence of the overall treatment-by-covariate interaction. In addition, a variable importance ranking feature is facilitated via random forests of interaction trees to help to determine important effect modifiers of the BCEI.

Tham2PIIIT3

Identifying Subgroups in Clinical Trials via Random Forests and Regression Trees

Jared Foster, Jeremy M.G. Taylor, Stephen J. Ruberg

University of Michigan, Ann Arbor, USA

We consider the problem of identifying a subgroup of patients who may have an enhanced treatment effect in a randomized clinical trial, and it is desirable that the subgroup be defined by a limited number of covariates. For this problem, the development of a standard, pre determined strategy may help to avoid the well-known dangers of subgroup analysis. We present a method developed to find subgroups of enhanced treatment effect. This method, referred to as "Virtual Twins", involves predicting response probabilities for treatment and control "twins" for each subject. The difference in these probabilities is then used as the outcome in a classification or regression tree, which can potentially include any set of the covariates. We define a measure Q(A^) to be the difference between the treatment effect in estimated subgroup A^ and the marginal treatment effect.We present several methods developed to obtain an estimate of Q(A^), including estimation of Q(A^) using estimated probabilities in the original data, using estimated probabilities in newly simulated data, two cross-validation-based approaches and a bootstrap-based bias corrected approach. Results of a simulation study indicate that the Virtual Twins method noticeably outperforms logistic regression with forward selection when a true subgroup of enhanced treatment effect exists. Generally, large sample sizes or strong enhanced treatment effects are needed for subgroup estimation. Additionally, simulation results suggest that the Virtual Twins method is fairly insensitive to moderate variations in the true model for the observations.

Tham2PIIIT4

$\mu$TOSS - Multiple hypotheses testing in an open software system

Wiebke Werft, Thorsten Dickhaus, Gilles Blanchard, Niklas Hack, Frank Konietschke, Kornelius Rohmeyer, Jonathan R.

German Cancer Research Center Heidelberg and Humboldt-University Berlin, Germany

Most of the research in the field of multiple hypotheses testing has immediate applications in the life sciences and, consequently, theoretically derived methods are typically more or less directly implemented into individual software. It is fair to say that up to now every research group uses its own implementations, making (simulation) study evaluations and related results not entirely comparable. Moreover, the spread of newly emerging methods is hindered by the lack of a common software platform to agree on.

Following a suggestion by Yoav Benjamini in his keynote talk at MCP 2009, we present an R-based, open software framework for multiple hypotheses testing called "$\mu$TOSS", sponsored by the PASCAL2 European Network of Excellence and realized at Berlin Institute of Technology in 2010. General key assets of the $\mu$TOSS system are:

- Source code-open implementation (using R)
- Well-documented developer interfaces for new procedures to add-on
- Graphical user interface ($\mu$TOSS GUI)
- Online user's guide on which procedure to use according to the user's specification of the test problem
- Inclusion of a large part of the known MCP methods
- Inclusion of testbed datasets for verification and exemplary purposes
- Ongoing maintenance (via R-Forge and C-RAN)

The several components of the $\mu$TOSS system provide
(i) multiple tests controlling the Family-Wise Error Rate (single-step and stepwise rejective methods, resampling-based procedures),
(ii) multiple tests controlling the False Discovery Rate (classical and data-adaptive frequentistic methods as well as Bayesian approaches and resampling-based techniques),
(iii) multiplicity-adjusted simultaneous confidence intervals,
and will be exemplified with real-life datasets.

Reference:
Gilles Blanchard, Thorsten Dickhaus, Niklas Hack, Frank Konietschke, Kornelius Rohmeyer, Jonathan Rosenblatt, Marsel Scheer, Wiebke Werft:
μTOSS - Multiple hypothesis testing in an open software system
Journal of Machine Learning Research: Workshop and Conference Proceedings, Vol. 11 (2010), 12-19.

Thpm1PIT1

gMCP - an R Package for Graphical Multiple Test Procedures

Kornelius Rohmeyer, Florian Klinglmueller

University of Hannover, Germany

The relations and priorities between elementary hypotheses in a multiple test problem often can be adequately described by a weighted graph as Bretz et al. (2009) and Burman et al. (2009) have shown.

With the open source R package gMCP we provide a framework and Java based graphical user interface to design appropriate graphs for test problems, perform corresponding Bonferroni-based or parametric tests and calculate compatible simultaneous confidence intervals as well as adjusted p-values.

The talk will show the usage of the package and explain helpful features with real-life examples from the literature.

References:
F. Bretz, W. Maurer, W. Brannath, M. Posch (2009). A graphical approach to sequentially rejective multiple test procedures.
Statistics in Medicine 2009; 28:586–604.
F. Bretz, M. Posch, E. Glimm, F. Klinglmueller, W. Maurer, and K. Rohmeyer (2011).
Graphical approaches for multiple comparison problems using weighted Bonferroni, Simes or parametric tests.
Biometrical Journal, to appear.
C.-F. Burman, C. Sonesson, O. Guilbaud. A recycling framework for the construction of Bonferroni-based multiple tests.
Statistics in Medicine 2009; 28:739-761
K. Rohmeyer, F. Klinglmueller (2011). gMCP: Graph Based Multiple Test Procedures
R package version 0.6-6. URL http://CRAN.R-project.org/package=gMCP/.

Thpm1PIT2

SiZ-MCP: A New Tool for Sample Size Calculations for MCPs

Cyrus Mehta, Lingyun Liu, Pralay Senchaudhuri, Yannis, Jemiai

Cytel Inc, USA

In this presentation we will demonstrate SiZ-MCP a new tool with a graphical user interface for simulation based sample size calculations for multiple comparison procedures that compare several treatments to a common control. The following procedures are provided: Dunnett single step, Dunnett step-down, Bonferroni, Sidak, Weighted Bonferroni, Holm, Hochberg, Hommel, Fixed Sequence and Fallback. To our knowledge, no such tool for sample size calculations currently exists. In the talk we will identify settings in which one method is more powerful than another and will make some recommendations for clinical trials.

Thpm1PIT3

New SAS Tools for Multiple Comparisons in Very General Models

Randy Tobias, Peter Westfall, Russ Wolfinger

SAS Institute Inc., USA

SAS/STAT software has many powerful procedures for fitting general models with complicated effects. Traditionally, extensive postprocessing facilities for a fitted model, including multiplicity adjusted comparisons between group LS-means, have been limited to a few procedures, such as GLM and MIXED. This presentation discusses new facilities in SAS that provide a full complement of multiple comparisons for a wide spectrum of models. These facilities are newly available within many procedures for specifying at the time of analysis, but you can also store a fitted model and restore it later to use with the new PLM procedure for post-fit analysis.

Thpm1PIT4

Consonance and the Closure Method in Multiple Testing

Michael Wolf, Joseph Romano, Azeem Shaikh

University of Zurich, Switzerland

Consider the problem of testing s hypotheses simultaneously. In order to deal with the multiplicity problem, the classical approach is to restrict attention to procedures that control the familywise error rate (FWE). Typically, it is known how to construct tests of the individual hypotheses, and the problem is how to combine them into a multiple testing procedure that controls the FWE. The closure method of Marcus et al. (1976), in fact, reduces the problem of constructing multiple test procedures which control the FWE to the construction of single tests which control the usual probability of a Type 1 error. The purpose of this paper is to examine the closure method with emphasis on the concepts of coherence and consonance. It was shown by Sonnemann and Finner (1988) that any incoherent procedure can be replaced by a coherent one which is at least as good. The main point of this paper is to show a similar result for dissonant and consonant procedures. We illustrate the idea of how a dissonant procedure can be strictly improved by a consonant procedure in the sense of increasing the probability of detecting a false null hypothesis while maintaining control of the FWE. We then show how consonance can be used in the construction of some optimal maximin procedures.

Thpm1PIIT1

A Consonant Partition Testing Strategy for Multiple Endpoints

Bushi Wang, Xinping Cui

Boehringer Ingelheim, USA

To evaluate efficacy in multiple endpoints in confirmatory clinical trials is a challenging problem in multiple hypotheses testing. The difficulty comes from the different importance of each endpoint and their underlying correlation. Current approaches to this problem are based on closed testing or partition testing, which test the efficacy in certain dose-endpoint combinations and collate the results. Partition testing is in general a more powerful approach since it tests fewer hypotheses to avoid unnecessary power loss. Despite their different formulations, all current approaches test their dose-endpoint combinations as intersection hypotheses and apply various union-intersection tests. Likelihood ratio test is seldomly used due to the extensive computation and lacks of consistent inferences.

In this article, we first generalize the decision path principle proposed by Liu and Hsu (2009) to the cases with alternative primary endpoints and co-primary endpoints. Then we propose a new partition testing approach which is based on consonance adjusted likelihood ratio test. The new procedure provides consistent inferences and yet it is still conservative and does not rely on the estimation of endpoint correlation or independence assumptions which might be challenged by regulatory agencies.

Thpm1PIIT2

An Interval Property for Multiple Testing Procedures

Harold Sackrowitz, Arthur Cohen

Rutgers University, USA

We begin with the realization that all multiple testing procedures, no matter how complex, do induce tests on the individual hypothesis testing problems under consideration. These individual induced tests are often quite complicated and rarely studied. We will see that there are some desirable monotonicity and convexity properties that these induced tests often lack. Implications of the lack of, what we refer to as the interval property, will be discussed. Also a method of constructing stepwise procedures that do have the property will be presented.

Wam1PIIT4

Joint Models and Tests for Time to Tumor Recurrence and Disease Stage in Oncology Clinical Trials

Olga Marchenko, Prof. R. Keener; Prof. A. Tsodikov

University of Michigan, Ann Arbor, USA

In this presentation, a clinical trial with bladder cancer patients who went through surgery and were followed up for tumor recurrence will be discussed. The surgery was conducted on patients with an early cancer stage. There was a control group using standard procedures and an experimental group with a drug designed to enhance observation of suspected cancer lesions. One of the primary objectives of the study was to evaluate and compare the time to tumor recurrence (or progression) of patients in the control and experimental groups. At the time of tumor recurrence, the disease stage was also evaluated. The majority of these stages were less advanced, while some patients progressed to more aggressive stages. The stage of the disease at recurrence significantly impacts future treatment and quality of life. Therefore, analyzing and comparing the time to tumor recurrence and the stage at recurrence jointly makes more sense than an analysis based primarily on the time to recurrence. This trial served as motivation for the current research.

Parametric and semi-parametric methods to model the joint distribution of recurrence stage and time to recurrence will be reviewed. Using these models, the methods to estimate and test treatment efficacy will be proposed.

Thpm1PIIT3

Alpha Maximized Multiplicity Adjustment in Genomic Studies using Sequential Post-Hoc Matching

Jimmy Efird

East Carolina Heart Institute, USA

Multiplicity adjustment poses a significant challenge in genomic association studies of disease risk.   Sequential Post-hoc matching is an efficient technique for increasing the number of SNPs examined in a fixed alpha setting. The sample size and power for this method is reviewed in this paper.

Thpm1PIIT4

Poster Session

Sample Size Calculation in Phase II Selection Designs

Zuoshun Zhang, Angela Hu

Celgene Corporation, USA

The statistical methods for ranking and selection of treatment arms were introduced and used in the designs of phase 2 oncology clinical trials, where subjects were randomized to several promising treatment arms with the goal to select one arm for further development. In recent years, the methods were generalized for survival endpoint and for different designs with binary endpoint. In order to facilitate its wider application, there need readily applicable methods for sample size calculations. In this note, we showed that the sample size can be calculated using exact binomial distribution for classical selection designs. For selection designs with survival endpoints, the design can be double blinded and it is desirable to follow the trial to a fixed number of events for all arms. By assuming exponential distribution of survival time and adapting Bechhofer's method on selection designs with normal endpoint, we developed a method to estimate sample size of total number of events for all arms and further verified the design have desirable correct selection probability using simulations. We proposed a new class of flexible designs with binary endpoints and gave an exact method calculating sample size with specified correct selection probability based on binomial distributions.

A Two Stage Procedure to Control the Generalized Family Wise Error Rate

Djalel Eddine Meskaldji, Jean-Philippe Thiran, Stephan Morgenthaler

Ecole Polythecnique Fédérale de Lausanne, Switzerland

The problem of multiple testing has generated a lot of discussion in many fields of research. An important question for the researcher related to the multiple testing is the choice of the control of the false discoveries occurrences. From the Bonferroni procedure, which controls strongly the Family Wise Error Rate (FWER), to the False Discovery Rate (FDR) control procedures, the researcher could choose a metric among a variety of measures that may be either stringent or relaxed according to the purpose of the study. As a modification of the FWER, Lehmann and Romano (LR) proposed a Bonferroni type procedure that uses k*alpha/m as a single test level and showed that the procedure controls the generalized-FWER, where m is the number of single tests. It is clear that k*alpha must be less than 1; otherwise, the control of the g-FWER poses no real restriction. In any cases, one can say that the (LR) procedure controls the Per Family Error rate at level k*alpha. We propose a new multiple testing procedure that controls the Per Family Error Rate (PFER) to be less than alpha+epsilon (epsilon<<alpha*(k-1)) in the weak sense and controls the PFER to be less than k*alpha in the strong sense. The procedure is adapted to spatial data (e.g. MRI images) but can be applied in the general case.

The proposed procedure works in two stages. First, we divide the family of tests into b blocks and we apply the Bonferroni procedure at level alpha/b for each single block test using the block's mean as a summary statistic of each block. As a second step of the proposed procedure, the LR procedure at level k*alpha is applied inside the significant blocks. We contrast the proposed procedure with some well-known alternatives in terms of power and expected number of false positives in independent and positive dependent cases.

On the Identification of Predictive Biomarkers in High-Dimensional Data

Wiebke Werft, Axel Benner

German Cancer Research Center, Germany

Our research has been motivated by a companion study to a randomized phase II neoadjuvant breast cancer trial. The primary objective of the companion study was to identify predictive genes for the response to two specific treatments. Prior to treatment, breast cancer tissue of the patients has been collected for microarray experiments. Specific genes should be identified that can potentially predict the patients response to treatment (i.e. presence or absence of pathological complete response in the breast) specifically to each of the two neoadjuvant regimens.

The identification of predictive biomarkers can be statistically addressed by inference of gene-wise generalised linear models (GLM) including an interaction term gene expression times treatment. Inference for such GLMs is then often based on likelihood-ratio or Wald test statistics to test the influence of interaction of gene expression and treatment on the clinical treatment response. For multiple testing scenarios coming along with these gene-wise GLMs the control of the false discovery rate (FDR) would be appropriate.

In a simulation study the utility of various FDR controlling multiple testing procedures for the identification of predictive genes is examined. Since the usual experiment on microarray data deals with small numbers of observations due to financial or probe limitations special interest lies on the sensitivity with respect to small sample sizes. Hence, different methods of inference on the interaction term were used to account for deficits due to small sample sizes. Regarding the correction for multiple testing, adaptive modifications of the Benjamini-Hochberg adjustment procedure were considered. To incorporate the dependency structure of the gene expression data and hence of the test statistics, the Benjamini-Yekutieli and the Blanchard-Roquain procedures were included in the analyses. Moreover, resampling-based joint MTPs also suitable for arbitrarily dependent test statistics were extended for the logistic regression model incorporating shift and scale-transformed and quantile-transformed joint null distributions for the step-down minP and maxT procedures.

The results of the simulation study reveal that sample size issues and the correct choice of test statistics together with an appropriate multiple testing procedure play a major role for controlling the FDR for the identification of predictive factors. Applications of methodologies to the motivating breast cancer study data correspond to the results of the simulation studies. Our research will provide guidance in establishing lists of potential predictive genes for usage in personalised medicine which will generally have less false positive detections.

Bayesian Testing for No Effect in Nonparametric Regression

Taeryon Choi, Jeongeun Kim

Korea University, South Korea

When we explain the response variable in terms of nonparametric regression model, we consider testing a null hypothesis that a predictor variable has no effect in regression analysis. For this purpose, we propose Bayesian model comparisons using Bayes factors for departures from constant mean in regression. The use of Bayes factor provides a Bayesian lack of fit testing for no effect and enables a correct model to be determined in the context of Bayesian model selection. We examine properties of Bayes factors for testing no effect under various situations. The theoretical validation of the Bayesian lack of fit procedure is investigated, and numerical illustrations for computing Bayes factors are also given with a synthetic data and a real data with application in trend detection.

Combining P-Values from Independent Studies

JaiWon Choi, Balgobin Nandram, Taeryon Choi

Medical College of Georgia, USA

When there are more than one p-value from independent studies for a same goal, we need to combine these p-values to report an overall result from these studies. One of the methods is the Fisher's score to combine such p-values by a chi-square transformation under the uniform assumption. However, the realized value for combined p-values often deviates from the nominal value of chi-square distribution when realized p-values are located lower or upper extremes of uniform distribution since they are correlated violating the assumption of independence. In this paper, we discuss the degree of such deviation, and propose how to adjust them properly dealing with correlation.

Multiple Testing Procedures with Applications to Whole-Genome Analysis

Hongmei Jiang

Northwestern University, USA

Multiple testing is a challenging problem in whole-genome studies where hundreds of thousands or millions of tests were performed simultaneously. For some array-based genomic experiments, such as copy number variation and methylation studies, the measurements of neighboring probes along the chromosome are highly correlations. Instead of testing each probe individually, we propose to perform the test on the genomic regions by taking into account of the correlation structure. Permutation-based technique is used to evaluate the statistical significance of the genomic regions. Both simulated and real data analyses will be used to compare the power of the proposed approach with existing methods.

Sample Size Determination in Clinical Trials with Two Correlated Co-Primary Time-to-Event Endpoints

Toshimitsu Hamasaki, Tomoyuki Sugimoto, Takashi Sozu

Osaka University Graduate School of Medcine, Japan

In cardiovascular or oncology clinical trials, two or three time-to-event variables may be investigated as co-primary endpoints, with the aim of providing a comprehensive picture of the treatment's benefits for subject's entire experience of a disease. For example, in oncology clinical trials, progression-free and overall survivals are frequently primary endpoints. In the case of more than one primary endpoint in clinical trials, there are two situations: one is to establish relevant benefits for all the primary endpoints, and other is to demonstrate any favorable benefit for at least one primary endpoint. This presentation will focus on the former situation and discuss sample size calculations for comparing the efficacy of two treatments on two co-primary time-to-event variables.

In this presentation, in order to derive the formula for sample size calculation, we first model such two possibly correlated (bivariate) time-to-event variables with given correlation structure by using three typical families of copula model (i.e., Clayton, Frank, and positive stable). Under the situation where the log-rank statistics or weighted log-rank statistics is used for the comparison of two groups, we formulate the correlation between the bivariate weighted log-rank statistics from bivariate time-to-event variables and then provide a computational method for the covariance matrix as a basis for the sample size calculations. Based on these results, we provide several methods to calculate the sample size required to compare two correlated co-primary time-to-event endpoints between the two groups. In addition, we extend the methods to a situation where one endpoints is time-to-event variable, but other is binary variable. We perform a simulation study to evaluate the performance of the methods and give numerical examples to illustrate the aspect of the methods.

Testing the Equality of Pairs of Mean Vectors and Simultaneous Confidence Intervals in Elliptical Distributions

Aya Shinozaki, Takashi Seo

Tokyo University of Science, Japan

Often the data from a repeated-measurements experiment may consist of groups or repetitions of the same response at different times. We consider testing the equality of pairs of mean vectors under the repeated-measurements data and the simultaneous confidence intervals in elliptical distributions. In order to test and construct the simultaneous confidence intervals, we derive upper percentiles of paired T-squared statistic by using asymptotic expansion procedure. Further we investigate the effects of non-normality on upper percentiles of the paired T-squared statistic. Finally the accuracy of approximation is investigated by Monte Carlo simulations for selected values of parameters. An example with a high dimensional data is also given to illustrate the paired T-squared test.

Key Words: asymptotic expansion; elliptical distribution; paired T-squared statistic; simultaneous confidence intervals.

On the Distributions of Some Test Statistics for Profile Analysis with Two-step Monotone Missing Data

Mizuki Onozawa, Takashi Seo

Tokyo University of Science, Japan

We consider one-sample or two-sample profile analysis when the data has two-step monotone missing observations. For one-sample profile analysis, to test for equality of means (the profile is flat), the test statistic based on the maximum likelihood estimators of mean vector and covariance matrix with two-step monotone missing data is proposed and its asymptotic null distribution is derived. For two-sample profile analysis, there are three hypotheses of interest in comparing the profiles of two samples: two profiles are parallel, two profiles are same level, and two profiles are flat. The test statistics and their asymptotic null distributions for the three hypotheses are also given. Simultaneous confidence intervals can be found by using the upper percentiles of these test statistics. When the data has not missing observations, the test statistics reduce to the usual test statistics given, for example, in Morrison (2005). The behavior of distributions for the test statistics is investigated by Monte Carlo simulations. An example is also given.

Tests for Two Mean Vectors and Simultaneous Confidence Intervals with Unequal Covariance Matrices in Two-step Monotone Missing Data

Tamae Kawasaki, Takashi Seo

Tokyo University of Science, Japan

We consider the tests for equality of two mean vectors and simultaneous confidence intervals with unequal covariance matrices when the data has two-step monotone pattern missing observations. For the case of equal covariance matrices, Hotelling's T-squared test is discussed by Seko, Kawasaki and Seo (2011). In this paper, by using the maximum likelihood estimators of mean vector and covariance matrix for two-step monotone missing data, we give Hotelling's T-squared type statistics for two-sample problem with unequal covariance matrices and propose their approximate upper percentiles. The simultaneous confidence intervals for all linear compounds of the difference of two mean vectors are also given. The accuracy of approximation to the upper percentiles of Hotelling's T-squared type statistic is investigated by Monte Carlo simulations for some selected parameters. An example is also given.

Presenters Index

Jiang	Hongmei	Poster Session	120
Kaizar	Eloise	Tupm2PIIIT1	44
Kawasaki	Tamae	Poster Session	124
Klinglmueller	Florian	Tupm1PI+IIT1	31
Koenig	Franz	Tupm2PI+IIT1	39
Kordzakhia	George	Tham1PIIT2	82
Lababidi	Samir	Tuam2PIIIT3	29
Lai	Yinglei	Wam2PIIIT4	64
Landwehr	Sandra	Tupm1PIIIT3	37
Li	Huajiang	Wam1PIIIT4	56
Lipkovich	Ilya	Tham2PIIIT2	100
Liu	Yi	Tuam2PI+IIT2	24
Liu	Lingyun	Tupm1PI+IIT2	32
Liu	Wei	Wam1PIIT1	48
Liu	Qing	Wpm1PIT3	67
Liu	Aiyi	Tham2PIT3	93
Lu	Tong-Yu	Tham2PIIT3	97
Luo	Xiaolong	Wpm1PIT2	66
Magirr	Dominic	Tupm2PI+IIT3	42
Marchenko	Olga	Thpm1PIIT3	111
Maurer	Willi	Tham1PIIT4	84
McLain	Alexander	Wpm1PIIT3	71
Mehrotra	Devan	Tham2PIIT1	95
Mehta	Cyrus	Thpm1PIT3	106
Meskaldji	Djalel Eddine	Poster Session	115
Mueller	Peter	Tupm1PIIIT1	35
Nishiyama	Takahiro	Wam2PIIT4	60
Onozawa	Mizuki	Poster Session	123
Posch	Martin	Tuam2PI+IIT3	25
Rauch	Geraldine	Tham1PIT2	78
Riou	Jérémie	Tham1PIIIT3	87
Rohmeyer	Kornelius	Thpm1PIT2	105
Roquain	Etienne	Wpm1PIIIT3	75
Rosenblum	Michael	Wam1PIIT3	47

Rosenkrantz	Walter	Wam1PIIT2	49
Roths	Scott	Wpm1PIIIT4	76
Ruberg	Steve	Tham2PIIIT1	99
Sackrowitz	Harold	Wam1PIIT4	110
Sarkar	Sanat	Wpm1PIIIT1	73
Scheer	Marsel	Wam1PIIIT3	54
Shinozaki	Aya	Poster Session	122
Shkedy	Ziv	Tuam2PIIIT2	28
Speed	Terry	Tuam1PBallroomT2	21
Strassburger	Klaus	Wam1PIIIT1	52
Su	Xiaogang	Tham2PIIIT3	101
Sugitani	Toshifumi	Tupm2PI+IIT2	41
Takahashi	Sho	Wam2PIIT3	59
Tamhane	Ajit	Tham2PIT2	92
Tang	Szu-Yu	Tham1PIIIT2	86
Tobias	Randy	Thpm1PIT4	107
Troendle	James	Tupm2PIIIT2	45
Tsong	Yi	Tham1PIT1	77
Wang	Sue-Jane	Tuam2PI+IIT4	26
Wang	Jinde	Wam2PIIIT2	72
Wang	Bushi	Thpm1PIIT2	109
Wen	Miin-Jye	Wpm1PIIT2	70
Werft	Wiebke	Poster Session	116
Werft/Dickhaus	Wiebke/Thorsten	Thpm1PIT1	103
Wiens	Brian	Tham1PIIT3	83
Wolf	Michael	Thpm1PIIT1	108
Xu	Jin	Tham1PIT4	80
Xu	Yunling	Tham2PIIT2	96
Zhang	Yu	Tupm1PIIIT4	38
Zhang	Zuoshun	Poster Session	114