art work by Charlotte Lu

Conference website: http://www.mcp-conference.org

The 10th International Conference on Multiple Comparison Procedures

June 20- June 23, 2017

University of California, Riverside

Imprint:

MCP 2017, Riverside, CA USA
10th International Conference on Multiple Comparison Procedures
June 20 – June 23, 2017

Production: Xinping Cui, Ajit C Tamhane, Chenwei Tian

Table of Contents

Sponsors of the MCP 2017 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

Cytel

MERCK
INVENTING FOR LIFE

Takeda

PROSOFT
Clinical

AMGEN® 35 YEARS

UC RIVERSIDE
UNIVERSITY OF CALIFORNIA
Department of Statistics
statistics.ucr.edu

Keynote Speaker

- Jason Hsu (The Ohio State University, USA)

Invited Speakers

- Marina Bogomolov (Technion, Israel Institute of Technology)
- Xinping Cui (University of California, Riverside)
- Yingying Fan (University of Southern California)
- Helmut Finner (Institute for Biometrics and Epidemiology, German Diabetes Center (DDZ), Leibniz)
- William Fithian (University of California, Berkeley)
- Wenge Guo (New Jersey Institute of Techonology)
- Thomas Jaki (Lancaster University)
- Adel Javanmard (University of Southern California)
- Florian Klinglmueller (AGES - Austrian Agency for Health & Food Safety)
- Gene Pennello (FDA)
- Koji Tsuda (University of Tokyo)
- Pantelis Vlachos (Cytel, Inc.)

General Information

Conference Venue:
Highlander Union Building (HUB)
University of California, Riverside
Riverside, CA, 92507

Conference Website:
www.mcp-conference.org

Important Dates

June 10 End online registration deadline
June 20 Short courses (8:30am – 12:30pm
1:30pm – 5:30pm)
June 20 Social Mixer (6:00pm – 9:00pm)
June 21 Start main conference
June 22 Conference excursion: beach excursion / dinner
June 23 Conference end

Scientific Program

Short Courses

TUESDAY, JUNE 20 8:30am – 12:30pm

8:00 am – 6:00 pm **CONFERENCE REGISTRATION** **HUB 3rd floor**

SHORT COURSES

TAM-1. Fundamentals of Multiple Testing and Graphical Approaches to Multiple Testing Problems — HUB 367
Haiyan Xu, Johnson & Johnson
Dong Xi, Novartis
Jason C. Hsu, The Ohio State University

TAM-2. Multiple Hypothesis Testing in Group Sequential and Adaptive Clinical Trials — HUB 379
Christopher Jennison, University of Bath

12:30 pm – 1:30 pm **LUNCH (ON YOUR OWN)** **HUB food court**

TUESDAY, JUNE 20 1:30pm – 5:30pm

SHORT COURSES

TPM-1. Trial Designs with Multiple Treatments and Multiple Endpoints Using East ® — HUB 367
Cyrus Mehta, Cytel Inc.
Lingyun Liu, Cytel Inc.

TPM-2. Artificial Intelligence, Machine Learning, and Precision Medicine — HUB 379
Haoda Fu, Eli Lilly

6:00pm-9:00 pm **SOCIAL MIXER** HUB 355

Sessions

WEDNESDAY, JUNE 21 8:30am – 10:00am

8:00 am – 6:00 pm **CONFERENCE REGISTRATION** **HUB 3rd floor**

KEYNOTE SESSION **HUB302**
Chair: Xinping Cui

WAM 1-1. Opening of the Conference
Xinping Cui, Professor & Chair, Department of Statistics, UCR

History of MCP Conferences
Ajit C. Tamhane, Professor of IEMS and Statistics, Northwestern University

Welcome Remarks
Cynthia K. Larive, Interim Provost & Executive Vice Chancellor, UCR

Welcome Remarks
Kathryn Uhrich, Dean of College of Natural and Agricultural Sciences, UCR

WAM 1-2. **KEYNOTE SPEECH**
Errors in Multiple Testing Big and Small, Now and Then, More or Less
Jason Hsu, The Ohio State University

10:00am-10:30 am **COFFEE BREAK** **HUB 3rd floor**

WEDNESDAY, JUNE 21 10:30am – 12:10pm

WAM2-1 PANEL DISCUSSION HUB 355

Organizers/Chairs: Xinping Cui/Jason Hsu

Panel on Multiplicity Issues in Clinical Trials
Martin Posch, Medical University of Vienna, Austria
Florian Klinglmueller, Medical University of Vienna, Austria
Bushi Wang, Boehringer Ingelheim, USA
Yuki Ando, Pharmaceuticals and Medical Devices Agency, Japan
Haiyan Xu, Johnson & Johnson, USA
Dong Xi, Novartis, USA

WAM2-2 High Dimensional and Large Scale Problems HUB 367

Organizer/Chair: Joseph Romano

10:30-11:00 Analysis of Error Control in Large Scale Two-Stage Multiple Hypothesis Testing (Invited Talk)
Wenge Guo*, Joseph Romano, New Jersey Institute of Technology, USA

11:00-11:20 AdaPT: An Interactive Procedure for Multiple Testing with Side Information
Lihua Lei*, William Fithian, University of California, Berkeley, USA

11:20-11:40 New Procedures Controlling the False Discovery Proportion via Romano-Wolf's Heuristic
Etienne Roquain*, Sylvain Delattre, Universite Pierre et Marie Curie, France

11:40-12:00 Bonferroni-Type Adjustments in MCPs for One-Sided Hypotheses
Michael Wolf*, University of Zurich, Switzerland

12:00-12:10 Floor Discussion

WAM2-3 Multiple Testing HUB 379

Contributed Papers
Chair: Jian Zhu

10:30-10:50 General Covering Principle to Address Multiplicity in Hypothesis Testing
Huajiang Li*, Hong Zhou, Avanir Pharmaceuticals, USA

10:50-11:10 A Closed Testing Procedure Based on Ordered Alternatives in Dose-Response Studies
Girish Aras*, Amgen, USA

11:10-11:30 Tests for the Positive Dependence Assumption of Simes' Inequality
Jiangtao Gou*, Hunter College, USA

11:30-11:50 Non-Consonant Rejections in Hommel's Procedure
Jelle Goeman*, Aldo Solari, Leiden University Medical Center, Netherlands

11:50-12:10 Post Hoc Inference Through Joint Familywise Error Rate Control
Pierre Neuvial*, Gilles Blanchard, Etienne Roquain, CNRS and Toulouse Mathematics Institute, France

12:10pm-1:30 pm	**LUNCH**	**HUB food court**

WEDNESDAY, JUNE 21 1:30pm – 3:10pm

WPM1-1 **Subgroup Analysis I** **HUB 355**

Organizers: Jason Hsu, Dong Xi

Chair: Jason Hsu

1:30-2:00	Biomarker Subgroup Testing, Misclassification, and Missing Data (Invited Talk) Gene Pennello*, Jingjing Ye, FDA, USA
2:00-2:20	Partitioning to Guarantee Subgroup Sensitive Inference in Personalized Medicine Szu-Yu Tang*, Ventana Medical Systems, Inc., USA
2:20-2:40	Subgroup Finding via Bayesian Additive Regression Trees Siva Sivaganesan*, University of Cincinnati, USA
2:40-3:00	Exploration of Heterogeneous Teatment Effects via Concave Fusion Shujie Ma*, University of California, Riverside, USA
3:00-3:10	Floor Discussion

WPM1-2 **High Dimensional Variable Selection and Multiple Testing** **HUB 367**

Organizer/Chair: Xinping Cui

1:30-2:00	Model-free Knockoffs for High-dimensional Controlled Variable Selection (Invited Talk) Yingying Fan*, University of Southern California, USA
2:00-2:20	Multilayer False Discovery Rate Control for Variable Selection Eugene Katsevich*, Chiara Sabatti, Stanford University, USA
2:20-2:40	Penalized Likelihood and Multiple Testing Harold Sackrowitz*, Arthur Cohen, John Kolassa, Rutgers University, USA
2:40-3:00	Assessing Variable Selection Uncertainty in Linear Models Aldo Solari*, Ningning Xu, Jelle Goeman, University of Milano-Bicocca, Italy
3:00-3:10	Floor Discussion

WPM1-3 **Multiple Endpoints** **HUB 379**

Contributed Papers

Chair: Dong Xi

1:30-1:50	A Gatekeeping Procedure to Test a Primary and a Secondary Endpoint in a Group Sequential Design with Multiple Interim Looks Ajit Tamhane*, Jiangtao Gou, Christopher Jennison, Cyrus Mehta, Teresa Curto, Northwestern University, USA
1:50-2:10	How to Evaluate Type II Error Rate with Multiple Endpoints Bushi Wang*, Naitee Ting, Boehringer Ingelheim, USA
2:10-2:30	Improved Testing Procedures for Group Sequential Trials with a Primary and a Secondary Endpoint Huiling Li*, Jianming Wang, Xiaolong Luo, Celegene, USA
2:30-2:50	On Simultaneous Tests of Superiority and Noninferiority of Multiple Endpoints in Clinical Trials Jie Chen*, Tze L. Lai, Merck Research Laboratories, USA
2:50-3:10	Testing Superiority When Noninferiority of the Same Endpoint is Assessed in a Multiple Comparison Procedure Scott Beattie*, Jiajun Liu, Pedro Lopez-Romero, Eli Lilly and Company, USA

3:10pm-3:30 pm	**COFFEE BREAK**	**HUB 3rd floor**

WEDNESDAY, JUNE 21 3:30pm – 5:10pm

WPM2-1 Multiple Testing for Sequential Data HUB 355

Organizer/ Chair: Jay Bartroff

3:30-4:00	Online Rules for Control of False Discovery Rate (Invited Talk) Adel Javanmard*, University of Southern California, USA
4:00-4:20	Sequential Testing of Multiple Hypotheses Under Arbitrary Joint Distributions Michael Hankin*, Jay Bartroff, University of Southern California, USA
4:20-4:40	Methods for Multiple Testing Error Control on Sequential Data Jay Bartroff*, University of Southern California, USA
4:40-5:00	Sequential Multiple Testing with Generalized Error Control: An Asymptotic Optimality Theory Yanglei Song*, Georgios Fellouris, University of Illinois at Urbana-Champaign, USA
5:00-5:10	Floor Discussion

WPM2-2 Pattern Minning HUB 367

Organizers: Frank Bretz/Toshimitsu Hamasaki
Chair: Toshimitsu Hamasaki

3:30-4:00	Statistical Pattern Mining: An Overview (Invited Talk) Koji Tsuda*, University of Tokyo, Japan
4:00-4:20	Selective Inference for Predictive Pattern Minning Ichiro Takeuchi*, Shinya Suzumura, Yuta Umezu, Koji Tsuda, Nagoya Institute of Technology, Japan
4:20-4:40	Accounting for a Categorical Covariate in Significant Pattern Mining Llinares Lopez Felipe*, Laetitia Papaxanthos, Dean Bodenham, Damian Roqueiro, Karsten B, ETH Zurich, Switzerland
4:40-5:00	Controlling Familywise Error When Rejecting at Most One Null Hypothesis Each From a Sequence of Sub-Families of Null Hypotheses Geoff Webb*, Mark van der Laan, Monash University, Australia
5:00-5:10	Floor Discussion

WPM2-3 Adaptive Designs I HUB 379

Contributed Papers
Chair: Christopher Jennison

3:30-3:50	Application of Frequentist Guidelines in Bayesian Adaptive Designs Jian Zhu*, Yi Liu, Takeda Pharmaceuticals, USA
3:50-4:10	Blinded Sample Size Re-Estimation in Three-Arm Trials with 'Gold Standard' Design Tobias Mütze*, Tim Friede, University Medical Center Gottingen, Germany
4:10-4:30	Optimized Adaptive Enrichment Designs for Clinical Trials with a Sensitive Subpopulation Martin Posch*, Thomas Ondra, Sebastian Jobjoernsson, Carl-Fredrik Burman, Franz Koenig, Nigel S, Medical University of Vienna, Austria
4:30-4:50	Optimal Adaptive Enrichment Trials Thomas Burnett*, University of Bath, UK
4:50-5:10	Multi-Armed/Bandit Testing with Online FDR Control Fanny Yang*, Aaditya Ramdas, Kevin Jamieson, Martin Wainwright, University of California, Berkeley, USA

5:10pm-5:30 pm **COFFEE BREAK** **HUB 3rd floor**

WEDNESDAY, JUNE 21 5:30pm –6:30pm

WPM3-1 POSTER SESSION HUB 367

Statistical Properties of Bernstein copulae with Applications in Multiple Testing
Andre Neumann*, Taras Bodnar, Dietmar Pfeifer, Thorsten Dickhaus, University of Bremen, Germany

Simultaneous Confidence Intervals for Pairwise Comparisons Among Mean Vectors With Monotone Missing Data
Ayaka Yagi*, Takashi Seo, Tokyo University of Science, Japan

Constructing Tests to Compare Two Proportions Whose Critical Regions Guarantee to be Barnard Convex Sets
Jose Juan Castro Alva*, Felix Almendra-Arao, Hortensia Josefina Reyes-Cervantes, Facultad de Ciencias Fisico Matematicas de la Benemerita Universidad Autonoma de, Mexico

A Frequency-Domain Model Selection Criterion for a (Dynamic) Factor Model
Natalia Sirotko-Sibirskaya*, University of Bremen, Germany

A Computer-Assisted Pap Smear Screening System Based on Automated Cell Nuclei Segmentation
Fang-Hsuan Cheng*, Nai-Ren Hsu, Chung Hua University, Taiwan, China

Control of False Discoveries in Grouped Hypothesis Testing for eQTL Data
Pratyaydipta Rudra*, Andrew Nobel, Fred A. Wright, University of Colorado Denver, Anschutz Medical Campus, USA

FDR Control for Dependent Chi-Square Goodness of Fit Tests
Melinda McCann*, Amy Wagler, Oklahoma State University, USA

Multiplicity Correction in Group-Sequential Oncology Trials Including Subgroup Analyses and Multiple Primary Endpoints
Agnes Balogh*, Bristol-Myers Squibb, USA

Online FDR Control with Decaying Memory and Weights
Fanny Yang*, Aaditya Ramdas, Martin Wainwright, Michael Jordan, University of California, Berkeley, USA

Comparison of Different Variable Selection Methods in a Special Situation
Ningning Xu* Leiden University Medical Center, Netherlands

Decentralized Decision Making on Networks with False Discovery rate Control
Jianbo Chen*, Aaditya Ramdas, Michael Jordan, Martin Wainwright, University of California, Berkeley, USA

Post Selection Inference with Kernels
Yuta Umezu*, Makato Yamada, Kenji Fukumizu, Ichiro Takeuchi, RIKEN, Japan

Interactive Accumulation Test: A Flexible Framework for Structural Multiple Testing.
Lihua Lei*, William Fithian, University of California, Berkeley, USA

THURSDAY, JUNE 22 8:30am – 10:10am

ThAM1-1 PANEL on Adaptation Based on Blinded Data HUB 355

Organizers: Martin Posch/ Florian Klinglmüller

Chair: Martin Posch

8:30-8:40	Keeping the Blind Blind Janet Wittes*, Statistics Collaborative, Inc. USA
8:40-8:50	Blinded Adaptations of Clinical Trials - How Blind Do We Have to Be? Ekkehard Glimm*, Novartis Pharma AG, Switzerland
8:50-9:00	Bayesian Sample Size Re-estimation Incorporating External Data Tobias Mütze*, Tim Friede University Medical Center Gottingen, Germany
9:00-9:10	Estimation Following Blinded Adaptation Michael Proschan*, NIAID, NIH, USA
9:10-9:20	Discussant Christopher Jennison*, University of Bath, UK
9:20-10:10	Panel Discussion

ThAM1-2 Discrete FWER/FDR Methodology HUB 367

Contributed Papers

Chair: Sanat Sarkar

8:30-8:50	A Modified Benjamini-Hochberg Procedure for Discrete Data Sebastian Doehler*, Guillermo Durand, Etienne Roquain, Darmstadt Univerity of Applied Sciences, Germany
8:50-9:10	Discrete FDR Method Increases Sensitivity of Statistical Tests on Microbiome Data Lingjing Jiang*, Amnon Amir, Ruth Heller, Ery Arias-Castro, Rob Knight, University of California, San Diego, USA
9:10-9:30	Use of a Discrete False Discovery Rate Method for Flagging Potential Safety Signals in Clinical Trials Joseph Heyse*, Merck Research Laboratories, USA
9:30-9:50	Procedures Controlling the FWER for Discrete Data Li He*, Joseph Heyse, Merck Research Laboratories, USA
9:50-10:10	Exact Approach for Post Hoc Analysis of a Chi-Squared Test Guogen Shan*, Shawn Gerstenberger, University of Nevada Las Vegas, USA

ThAM1-3 Permutation/Resampling/Pattern Mining HUB 379

Contributed Papers

Chair: Jason Hsu

8:30-8:50	Implementing Monte Carlo Tests with Multiple Thresholds Georg Hahn*, Axel Gandy, Dong Ding, Imperial College London, UK
8:50-9:10	Estimating the Proportion of True Null Hypotheses Under Dependency Thorsten Dickhaus*, Andre Neumann, Taras Bodnar, University of Bremen, Germany
9:10-9:30	Permutation-Based Simultaneous Confidence Bounds for the False Discovery Proportion Jesse Hemerik*, Aldo Solari, Jelle Goeman, Leiden University Medical Center, Netherlands
9:30-9:50	Significant Pattern Mining on Graphs Mahito Sugiyama*, Osaka University, Japan
9:50-10:10	Controlling FWER and FDR in Emerging Pattern Mining Junpei Komiyama*, Masakazu Ishihata, Hiroki Arimura, Takashi Nishibayashi, Shinichi Minato, University of Tokyo, Japan

10:10am-10:30 am **COFFEE BREAK** **HUB 3rd floor**

THURSDAY, JUNE 22 10:30am – 12:10pm

ThAM2-1 Adaptive Designs II HUB 355

Organizer/Chair: Christopher Jennsion

10:30-11:00	Nonparametric Inference Following Adaptive Designs with Sample Size Reassessment (Invited Talk) Florian Klinglmueller*, Martin Posch, Livio Finos, AGES-Austrian Agency for Health & Food Safety, Austria
11:00-11:20	Correcting for Selection Bias in Adaptive Two-Stage Designs David Robertson*, Toby Prevost, Jack Bowden, MRC Biostatistics Unit, University of Cambridge, UK
11:20-11:40	Design and Monitoring of Multi-Arm Multi-Stage Clinical Trials Pranab Ghosh*, Cyrus Mehta, Boston University, USA
11:40-12:00	Analytical and Empirical Comparison of MAMS and P-Value Combination Approaches for Adaptive Designs Cyrus Mehta*, Pranab Ghosh, Lingyun Liu, Cytel Inc. USA
12:00-12:10	Floor Discussion

ThAM2-2 Weighted Multiple Testing Procedures HUB 367

Contributed Papers

Chair: Jian Zhu

10:30-10:50	Optimal Data-Driven Weighting Procedure with Grouped Hypotheses Guillermo Durand*, Universite Pierre et Marie Curie, France
10:50-11:10	A General Convex Framework for Multiple Testing with Prior Information Edgar Dobriban*, University of Pennsylvania, USA
11:10-11:30	A Unified Framework for Weighted Parametric Multiple Test Procedures Dong Xi*, Ekkehard Glimm, Willi Maurer, Frank Bretz, Novartis, USA
11:30-11:50	Conditionalized Testing: Improvement of Multiple Testing Methods When Testing Inflated p-Values Jakub Pecanka*, Jules Ellis, Jelle Goeman, Leiden University Medical Center, Netherlands
11:50-12:10	Adaptive Multiple Hypothesis Testing for Complex Networks and High Dimensional Data Djalel-Eddine Meskaldji*, Ecole polytechnique federale de Lausanne EPFL, Switzerland

ThAM2-3 FDR Methodologies HUB 379

Organizer/Chair: Sanat Sarkar

10:30-11:00	Selective Inference on a Tree of Hypotheses: New Error Rates and Controlling Strategies (Invited Talk) Marina Bogomolov*, Christine Burns Peterson, Yoav Benjamini, Chiara Sabatti, Technion, Israel Institute of Technology, Israel
11:00-11:20	Private False Discovery Rate Control and Robustness of the Benjamini Hochberg Procedure Weijie Su*, Cynthia Dwork, Li Zhang, University of Pennsylvania, USA
11:20-11:40	A Unified Treatment of Multiple Testing with Prior Knowledge Aaditya Ramdas*, Rina F. Barber, Michael I. Jordan, Martin J. Wainwright, University of California, Berkeley, USA
11:40-12:00	Optimal Rates and Tradeoffs in Multiple Testing (Canceled) Maxim Rabinovich*, Aaditya Ramdas, Martin Wainwright, Michael Jordan, University of California, Berkeley, USA
12:00-12:10	Floor Discussion

12:10pm-12:30 pm **Boxed Lunch** **HUB Food Court**

THURSDAY, JUNE 22 12:30pm – 2:00pm

ThPM1-1 21st Century Medicine: from Drugs to Biologics, from Generics to Biosimilars HUB 355

Organizer: Lingyun Liu/Dong Xi
Chair: Lingyun Liu

12:30-1:00	Flexible Statistical Approaches for Biosimilar Development (Invited Talk) Pantelis Vlachos*, Cytel Inc., USA
1:00-1:20	Unblinded Sample Size Re-Estimation in Bioequivalence Trials with Small Sample Sizes Sam Hsiao*, Lingyun Liu, Romeo Maciuca, Cytel Inc., USA
1:20-1:40	Understanding Biosimilarity by Totality of the Evidence: What should Statisticians Know Yushi Liu*, Jason Hsu, Eli Lilly and Company, USA
1:40-2:00	A Distribution-Free Consistency Adjusted Stepwise Testing Procedure (Canceled) Jaclyn McTague*, Dror Rom, Prosoft Clinical, USA

ThPM1-2 Subgroup Analysis II HUB 367

Organizer: Martin Posch
Chair: Jason Hsu

12:30-1:00	Confidence Regions for Treatment Effects in Biomarker Stratified Designs (Invited Talk) Thomas Jaki*, Fang Wan, Cornelia Kunz, Lancaster University, UK
1:00-1:20	Confident Inference for SNP Effects on Treatment Efficacy Ying Ding*, University of Pittsburgh, USA
1:20-1:40	Statistical Issues in Subgroup Discovery Using Permutation Testing Siyoen Kil*, LSK Global PS, Korea
1:40-2:00	A Case Study in Precision Medicine: Rilpivirine Versus Efavirenz for Treatment-Naive HIV Patients Zhiwei Zhang*, Wei Liu, Lei Nie, Guoxing Soon, UC Riverside, USA

ThPM1-3 Multiplicity or Bias in Biomarker Panel HUB 379

Organizer: Sue-Jane Wang/ Toshimitsu Hamasaki
Chair: Toshimitsu Hamasaki

12:30-1:00	Sequential Multiple Testing for Biomarker Discovery (Invited Talk) Xinping Cui*, Hailu Chen, University of California, Riverside, USA
1:00- 1:20	Unbiased Estimation of Biomarker Panel Performance When Combining Training and Testing Data in a Group Sequential Design Nabihah Tayob*, Kim-Anh Do, Ziding Feng, MD Anderson Cancer Center, USA
1:20-1:40	Elucidating Issues of Multiplicity that Arise with Clinical Trial Designs for Precision Medicine Brian Hobbs*, Nan Chen, MD Anderson Cancer Center, USA
1:40-2:00	An Application of FDR to Billions of Hypothesis Testing to Identify Expression Quantitative Trait Loci in Genome Wide Association Studies Irina Dinu*, Fahimeh Moradi, Elham Khodayari-Moez, University of Alberta, Canada

2:00pm-3:30 pm **SHUTTLE TO BEACH**

3:30pm-9:00 pm **BEACH EXCURSION/BEACH FRONT CONFERENCE DINNER**

FRIDAY, JUNE 23 8:30am – 10:10am

FAM1-1 Post-Selection Inference and Selective Inference HUB355

Organizer: Robert Tibishirani/ Jonathan Taylor
Chair: Xinping Cui

8:30-9:00	Adaptive Sequential Model Selection (Invited Talk) William Fithian*, Jonathan Taylor, Robert Tibshirani, Ryan Tibshirani, University of California, Berkeley, USA
9:00-9:20	Bootstrap Inference After Using Multiple Queries for Model Selection Jelena Markovic*, Jonathan Taylor, Stanford University, USA
9:20-9:40	Bayesian Post-Selection Inference in the Linear Model Snigdha Panigrahi*, Asaf Weinstein, Stanford University, USA
9:40-10:00	Selective Sign-Determining Multiple Confidence Intervals with FCR Control Asaf Weinstein*, Daniel Yekutieli, Stanford University, USA
10:00-10:10	Floor Discussion

FAM1-2 Recent Advances in Design and Analysis of Clinical Trials HUB367

Contributed Papers
Chair: Florian Klingmüller

8:30-8:50	Group Sequential Designs in Clinical trials with semi-competing risks outcomes Toshimitsu Hamasaki*, Koko Asakura, Scott R Evans, Tomoyuki Sugimoto, National Cerebral and Cardiovascular Center, Japan
8:50-9:10	Analysing Multiple Outcomes in Randomised Controlled Trials Using the Multilevel Multivariate Model Victoria Vickersta*, Gareth Ambler, Rumana Z Omar, University College London, UK
9:10-9:30	Comparison of Novel Approaches in Dose Response Studies Saswati Saha*, University of Bremen, Germany
9:30-9:50	Comparisons of Efficiency and Robustness of Multiple Testing Procedures in Phase 3 Clinical Trials Michael Lee*, Anjun Cao, Janssen R&D, USA
9:50-10:10	Testing Strategy in Phase 3 Trials with Multiple Doses David Li*, Pfizer, USA

FAM1-3 Genomics/Bioinformatics HUB379

Contributed Papers
Chair: Bushi Wang

8:30-8:50	Adaptive Filtering Multiple Testing Procedures for Partial Conjunction Hypotheses Jingshu Wang*, Art B. Owen, Chiara Sabatti, University of Pennsylvania, USA
8:50-9:10	A Novel FWER Controlling Procedure for Data with Reduced Rank Correlation Structure Xing Qiu*, University of Rochester, USA
9:10-9:30	An Empirical Bayes Test for Allelic-Imbalance Detection in ChIP-seq Qi Zhang*, Sunduz Keles, University of Nebraska Lincoln, USA
9:30-9:50	FDR Control on Directed Acyclic Graphs Jianbo Chen*, Aaditya Ramdas, , Michael Jordan, Martin Wainwright, University of California, Berkeley
9:50-10:10	Statistical Analysis for Estimating Multiple Stopped States in Walking Motions Toshinari Kamakura*, Kosuke Okusa, Chuo University, Japan

10:10am-10:30 am **COFFEE BREAK** **HUB 3rd floor**

FRIDAY, JUNE 23 10:30am – 12:10pm

FAM2-1 Decision Theory (Going Beyond FWER) HUB355

Organizer: Frank Bretz / Jason Hsu
Chair: Jason Hsu

10:30-11:00	From Higher Criticism and Local Levels of GOF Tests to Confidence Bounds for the Proportion of True Nulls (Invited Talk) Helmut Finner*, Veronika Gontscharuk, Klaus Strassburger, Institute for Biometrics and Epidemiology, German Diabetes Center (DDZ), Leibniz, Germany
11:00-11:20	Conditional Error Rate of Decision Made on the Secondary Endpoint Haiyan Xu*, Jason Hsu, Johnson & Johnson, USA
11:20-11:40	Optimal Statistical Decision for Gaussian Graphical Model Selection Petr Koldanov*, Alexander Koldanov, Valery Kalyagin, Panos Pardalos, NRU Higher School of Economics, Russia
11:40-12:00	Rank Verification for Exponential Families Kenneth Hung*, William Fithian, University of California, Berkeley, USA
12:00-12:10	Floor Discussion

FAM2-2 Recent Advance on Design and Analysis of Multi-regional Clinical Trials HUB367

Organizer: Toshimitsu Hamasaki / Chin-Fu Hsiao
Chair: Toshimitsu Hamasaki

10:30-10:55	Use of Interval Estimations in Design and Evaluation of Multi-Regional Clinical Trials Chin-Fu Hsiao*, Chieh Chiang, H.M. James Hung, Institute of Population Health Sciences, National Health Research Institutes, Taiwan
10:55-11:20	Multi-regional Biosimilarity Studies Victoria Chang*, Qi Xia, Boehringer-Ingelheim Pharmaceuticals Inc., USA
11:20-11:45	MRCT design models and drop-min data analysis. Fei Chen*, K. K. Gordon Lan, Gang Li, Janssen R&D, USA
11:45-12:10	Discussant Yuki Ando, PMDA, Japan

FAM2-3 Multiplicity Issues HUB379

Contributed Papers
Chair: Ramaiyan Elangovan

10:30-10:50	Comparing Several Variances with Control Using Sample Quasi Range Rajvir Singh*, Parminder Singh, Thapar University, India
10:50-11:10	Revisiting "What's Wrong with Bonferroni Adjustments" Andrew V. Frane*, University of California, Los Angeles, USA
11:10-11:30	The Reliability of Two Meta-Analysis Studies Stan Young*, Cheng You, CGStat, USA
11:30-11:50	Simultaneous Rank Tests for Pairwise Comparisons in Analysis of Covariance Hossein Mansouri*, Fangyuan Zhang, Texas Tech University, USA
11:50-12:10	Adaptive Designs in Clinical Trials Ramaiyan Elangovan*, Annamalai Univarsity, India

End of the Conference

12:10pm-1:30 pm	**LUNCH**	**HUB food court**

ABSTRACTS

TAM-1

Short Course
Fundamentals of Multiple Testing and Graphical Approaches to Multiple Testing Problems

Haiyan Xu , Johnson & Johnson
Dong Xi, Novartis
Jason C. Hsu, The Ohio State University

Two main principles provide the foundation of multiple testing: Closed testing and partitioning. Most multiple comparison methods can be derived and their validity can be proven using these two principles. In this course we show how they are connected using several examples. Starting with realistic numerical examples, the first and conceptual part of this short course will show that the traditional methods of Holm, Hochberg, and Hommel are special cases of closed testing and partitioning. To give insight into how the partitioning principle simplifies challenging problems, we show how Hsu and Berger (1999) formulated the problem of testing multiple doses in a pre-determined step-wise fashion to guarantee decision-making following a pre-specified path. We then show how Liu and Hsu (2009) applied the same path partitioning principle to simplify testing with multiple paths, such as testing for efficacy in multiple doses in combination with multiple endpoints. To conclude the first part of the course, we show how the gatekeeping method of Xu et al (2009), the graphical approach of Bretz et al (2011), and the partition testing principle of Liu and Hsu (2009) coincide and rely on the same testing principles.

The second part of this short course will be on the graphical approach's flexible and transparent implementation of multiple testing. Using graphical approaches (Bretz et al, 2009), one can easily construct and explore different test strategies and thus tailor the test procedure to the given study objectives. The resulting multiple test procedures are represented by directed, weighted graphs, where each node corresponds to an elementary hypothesis, together with a simple algorithm to generate such graphs while sequentially testing the individual hypotheses. We also present one case study to illustrate how the approach can be used in clinical practice. The presented methods will be illustrated using the graphical user interface from the gMCP package in R, which is freely available on CRAN.

Reference

1 Hsu, Jason C. and Berger, Roger L. (1999). Stepwise Confidence Intervals without Multiplicity Adjustment for Dose-Response and Toxicity Studies. Journal of the American Statistical Association, 94: 468-482.

2 Liu, Yi and Hsu, Jason C. (2009). Testing for efficacy in primary and secondary endpoints by partitioning decision paths. Journal of the American Statistical Association, 104: 1661-1670.

3 Xu, Haiyan and Nuamah, Isaac and Liu, Jingyi and Lim, Pilar and Sampson, Allan. (2009). A Dunnett-Bonferroni-based parallel gatekeeping procedure for dose-response clinical trials with multiple endpoints. Pharmaceutical statistics, 8: 301-316.

4 Bretz, Frank and Posch, Martin and Glimm, Ekkehard and Klinglmueller, Florian and Maurer, Willi and Rohmeyer, Kornelius. (2011). Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. Biometrical Journal, 53: 894-913.

5 Bretz, Frank and Maurer, Willi and Brannath, Werner and Posch, Martin. (2009). A graphical approach to sequentially rejective multiple test procedures. Statistics in medicine, 28: 586-604.

Email: hxu22@its.jnj.com, dong.xi@novartis.com, jch@stat.osu.edu

TAM-2

Short Course
Multiple Hypothesis Testing in Group Sequential and Adaptive Clinical Trials

Christopher Jennison, University of Bath

The course will introduce group sequential designs and their applications, including error-spending tests and inference for a secondary endpoint on termination of a group sequential trial. The complexity of the testing problem increases as more null hypotheses are tested during or after a sequential trial: we shall describe the general framework for such trials that combines the graphical approach to multiple hypothesis testing with group sequential tests of individual hypotheses. Adaptive designs allow mid-course modification of a trial while still protecting the type I error rate. Possible modifications include: enrichment designs, which shift their focus to a subset of the initial study population; seamless designs, which combine treatment selection and testing in a single trial; multi-arm Phase III trials which may drop treatments at interim analyses or stop early for a positive outcome. We shall describe the general approach to creating adaptive designs by combining close testing procedures and combination tests, and illustrate these ideas in a case study of a Phase III trial with treatment selection and a survival endpoint

Email: C.Jennison@bath.ac.uk,

TPM-1

Short Course
Trial Designs with Multiple Treatments and Multiple Endpoints Using East®

Cyrus Mehta, Cytel Inc.
Lingyun Liu, Cytel Inc.

Modern clinical trials are often designed to address multiple clinical questions which need multiplicity adjustments to ensure strong type I error control. Commonly encountered sources of multiplicities include multiple treatments, multiple endpoints, interim analyses and subgroup analyses. This workshop will cover three types of trial designs: (1) testing multiple endpoints with gatekeeping procedures, (2) compare multiple treatments/doses to a common control in group sequential design (MAMS), (3) seamless adaptive design with treatment selection and sample size re-estimation using p-value combination approach. These methods will be illustrated with the help of the East software with real clinical trial examples.

Emai: Lingyun.Liu@cytel.com, Cyrus.Mehta@cytel.com

TPM-2

Short Course
Artificial Intelligence, Machine Learning, and Precision Medicine

Haoda Fu, Eli Lilly

This half-day short course will provide an overview of statistical machine learning, and artificial intelligence techniques with applications to the precision medicine, in particular to deriving optimal individualized treatment strategies for personalized medicine. This short course will cover both treatment selection and treatment transition. The treatment selection framework is based on outcome weighted classification. We will cover logistic regression, support vector machine (SVM), ψ-learning, robust SVM, and angle based classifiers for multi-category learning, and we will show how to modify these classification methods into outcome weighted learning algorithms for personalized medicine. The second part of short course will also cover the treatment transition. We will provide an introduction on reinforcement learning techniques. Algorithms, including dynamic programming for Markov Decision Process, temporal difference learning, SARSA, Q-Learning algorithms, actor-critic methods, will be covered. We will discuss on how to use these methods for developing optimal treatment transition strategies. The techniques discussed will be demonstrated in R. This course is intended for graduate students who have some knowledge of statistics and want to be introduced to statistical machine learning, or practioners who would like to apply statistical machine learning techniques to their problems on personalized medicine and other biomedical applications.

Email: fu_haoda@lilly.com

WAM1-2

Keynote Speech
Errors in Multiple Testing Big and Small, Now and Then, More or Less

Jason Hsu, The Ohio State University, USA

All things are connected, old and new, here and there, this and that. For example, correcting error in bias from marginal means multiple comparisons in linear models by switching to least squares means in the early 1990s does not readily extend to binary or time-to-event outcomes, because odd ratios and hazard ratios are not be subgroup mixable. Therefore, multiple comparison methods have been developed recently for relative response and ratio of medians which are subgroup mixable, a fundamental requirement for logical inference in personalized medicine. As another example, a typical genome-wide association studies (GWAS) assesses whether any of a large number of single-nucleotide polymorphisms (SNPs) affect the phenotype. Because the statistical information in a SNP is small, turns out effect in any causal SNP makes all the zero-null no-association hypotheses false, so Type I error rate control becomes difficult to interpret. However, Tukey's (1954) non-coverage error rate control per confidence interval family has no such issue, and simultaneous confidence intervals for assessing SNP effects on clinical efficacy have recently been developed accordingly. In the GWAS setting, there are frequentists and Bayesians error rate considerations, both conditional and unconditional, co-developed by statistical learners and multiple testers. Such considerations can extend beyond GWAS. All Things Are Connected!

Email: jch@stat.osu.edu

WAM2-1

Panel Discussion
Multiplicity Issues in Clinical Trials

Martin Posch, Medical University of Vienna, Austria
Florian Klinglmueller, Medical University of Vienna, Austria
Bushi Wang, Boehringer Ingelheim, USA
Yuki Ando, Pharmaceuticals and Medical Devices Agency, Japan
Haiyan Xu, Johnson & Johnson, USA
Dong Xi, Novartis, USA

The panel discussion will provide general reflection on FDA's Multiple Endpoints Guidance and EMA Multiplicity Guidance. The panel discussion will also provide forward-looking considerations on the guidelines for guidance toward (known) emerging problems as well as principles toward potential (unanticipated) future problems.

WAM2-2 T1

Analysis of Error Control in Large Scale Two-Stage Multiple Hypothesis Testing

Wenge Guo, Joseph Romano
New Jersey Institute of Technology, USA

When dealing with the problem of simultaneously testing a large number of null hypotheses, a natural testing strategy is to first reduce the number of tested hypotheses by doing screening or selection, and then to simultaneously test selected hypotheses. The main advantage of this strategy is to greatly reduce the severe effect of high dimensions. However, the first screening or selection stage must be properly accounted for in order to maintain some type of error control. In this talk, we will introduce a selection rule based on the selection statistic which is independent of the test statistic when the tested hypothesis is true. Combining this selection rule and the conventional Bonferroni procedure, we can develop a powerful and valid two-stage procedure. The suggested procedure has several nice properties: (i) completely remove the selection effect; (ii) reduce the multiplicity effect; (iii) do not waste any samples while carrying out both selection and testing. Asymptotic power analysis and simulation studies illustrate that this proposed method provides higher power compared to usual multiple testing methods while controlling the type 1 error rate. Optimal selection thresholds are also derived based on our asymptotic analysis.

Email: wenge.guo@njit.edu

WAM2-2 T2

AdaPT: An Interactive Procedure for Multiple Testing with Side Information

Lihua Lei, William Fithian
University of California, Berkeley, USA

We consider the problem of multiple hypothesis testing with generic side information: for each hypothesis Hi we observe both a p-value pi and some predictor xi encoding contextual information about the hypothesis. For large-scale problems, adaptively focusing power on the more promising hypotheses (those more likely to yield discoveries) can lead to much more powerful multiple testing procedures. We propose a general iterative framework for this problem, called the Adaptive p-value Thresholding (AdaPT) procedure, which adaptively estimates a Bayes-optimal p-value rejection threshold and controls the false discovery rate (FDR) in finite samples. At each iteration of the procedure, the analyst proposes a rejection threshold and observes partially censored p-values, estimates the false discovery proportion (FDP) below the threshold, and either stops to reject or proposes another threshold, until the estimated FDP is below α. Our procedure is adaptive in an unusually strong sense, permitting the analyst to use any statistical or machine learning method she chooses to estimate the optimal threshold, and to switch between different models at each iteration as information accrues.

Email: lihua.lei@berkeley.edu

WAM2-2 T3

New Procedures Controlling the False Discovery Proportion via Romano-Wolf's Heuristic

Etienne Roquain, Sylvain Delattre
Universite Pierre et Marie Curie, France

Romano and Wolf (2007) have proposed a general principle that builds false discovery proportion (FDP) controlling procedures from k-family-wise error rate controlling procedures while incorporating dependencies in an appropriate manner. In this talk, we provide a careful theoretical study of this heuristic. This results in new methods overcoming the existing procedures with a proven FDP control.

Email: etienne.roquain@upmc.fr

WAM2-2 T4

Bonferroni-Type Adjustments in MCPs for One-Sided Hypotheses

Michael Wolf
University of Zurich, Switzerland

We consider a setting where the individual hypothesis all (i) concern univariate parameters and are (ii) one-sided. In such a setting, power gains can be obtained if adjustments to the global null sampling distribution are made for hypotheses that are "deep" in the individual null. Such adjustments are generally ad hoc and motivated on asymptotic grounds; in particular, they do not necessarily guarantee good finite-sample control of the family-wise error rate (or other error rates). In this talk, we will consider a Bonferroni-type adjustment that is motivated by finite-sample considerations instead. The performance of the method will be compared to the performance of previous proposals by means of Monte Carlo simulations.

Email: michael.wolf@econ.uzh.ch

WAM2-3 T1

General Covering Principle to Address Multiplicity in Hypothesis Testing

Huajiang Li, Hong Zhou
Avanir Pharmaceuticals, USA

Recently the covering principle was proposed to address the multiplicity issue when the decision order of testing individual null hypotheses implied coverage relations and constraints in multiple testing problems. The essence of the covering principle is based on the sample space partitioning instead of the parameter space partitioning as classical closed testing and partitioning principle did. Our current research extended the covering principle to a very general form that can simultaneously deal with multiple coverage relations in testing individual null hypotheses. We proposed a concept called maximum constrained class and decomposed the whole family of individual null hypotheses into a few overlapped sub-families. We proved that the multiple testing procedure constructed using the general covering principle strongly controls the familywise error rate as long as the multiple tests for each sub-familiy strongly control the type I error. Several examples from clinical trials were provided for the illustration purpose.

Email: hli@avanir.com

WAM2-3 T2

A Closed Testing Procedure Based on Ordered Alternatives in Dose-Response Studies

Girish Aras
Amgen, USA

In dose-finding trials, clinicians may believe a priori that the ordering of the responses is known, so the efficacy is assumed to increase monotonically with dose. Under order restricted alternatives, the likelihood ratio tests have far greater power than omnibus procedures, and in addition, they provide protection against the possibility of specifying an incorrect functional form. To establish a dose-response profile, not only we need to test doses against the placebo group, but also doses among themselves. With multiple hypotheses to be tested, protecting family-wise error rate becomes an issue. A closed testing principle was formulated by Marcus, Peritz and Gabriel1 and has been the mathematical foundation for multiple testing procedures. In general, we need to consider all possible intersections of the null hypothesis of interest. A hypothesis is rejected if its associated test and all tests associated with hypotheses implying it are significant. Order restricted inference combined with a closed testing procedure offers a neat solution to establish a dose-response profile that protects family-wise error rate in a strong sense.

Email: garas@amgen.com

WAM2-3 T3

Tests for the Positive Dependence Assumption of Simes' Inequality

Jiangtao Gou
Hunter College, USA

The commonly used familywise-error-rate-controlling procedures (Hochberg, 1988; Hommel 1988) and false-discovery-rate-controlling procedures (Benjamini and Hochberg, 1995) are based on the Simes (1986) test. Sarkar and Chang (1997) and Sarkar (1998) proved that the Simes test is conservative under a certain type of positive dependence. Later the assumption of the positive dependence is slightly relaxed. Gou and Tamhane (2017) provided an example to show that this assumption cannot be further relaxed to the positive quadrant dependence. Hence a weak assumption of positive dependence may not guarantee the conservativeness of the Simes test. In this presentation, we proposed tests specifically for the positive dependence assumption of the Simes test.

Email: jgou@u.northwestern.edu

WAM2-3 T4

Non-Consonant Rejections in Hommel's Procedure

Jelle Goeman, Aldo Solari
Leiden University Medical Center, Netherlands

The combination of closed testing with Simes local tests is exploited in Hommel's and Hochberg's famous procedures. Such closed testing procedures have non-consonant rejections, i.e. rejections of intersection hypotheses that are not implied by rejected elementary hypotheses. Earlier, we have shown that such non-consonant rejections can be exploited to generate simultaneous confidence bounds for the false discovery proportion for every subset of the hypotheses. These bounds can be more powerful than those arising naively from the results of Hommel's method. In this talk we investigate the scalability of this closed testing procedure as the number of hypotheses goes to infinity in a finite sample, using a general formulation of Efron's empirical Bayes model. We show that the proportion of rejected elementary hypotheses among the false hypotheses converges to zero, illustrating the lack of scalability of familywise error rate. However, we also show that, as the number of hypotheses goes to infinity, for every desired false discovery proportion $q>0$, and every confidence level $1-\alpha$, there is a set for which we can establish $1-\alpha$ confidence of a false discovery proportion at most q. If a minimal amount of signal is present, the size of this set grows linearly with the number of hypotheses. This demonstrates that non-consonant rejections, which are rare in closed testing with Simes local tests in small problems, are ubiquitous in larger problems. It also demonstrates that, unlike familywise error statements, false discovery proportions obtained from closed testing procedures do scale well with the size of the multiple testing problem.

Email: j.j.goeman@lumc.nl

WAM2-3 T5

Post Hoc Inference Through Joint Familywise Error Rate Control

Pierre Neuvial, Gilles Blanchard, Etienne Roquain
CNRS and Toulouse Mathematics Institute, France

The objective of post hoc inference in a multiple testing context is to devise procedures able to provide a statistical guarantee on any candidate set of rejected hypotheses, including user-defined and/or data-driven candidate sets. We introduce a general methodology for post hoc inference. This methodology relies on the control of a multiple testing error rate that we call the joint Family-Wise Error Rate (JR). Our construction generalizes existing post hoc procedures under positive dependence proposed by Goeman and Solari (Statistical Science, 2011). We propose a generic approach to build JR-controlling procedures in the situation where the joint null distribution of the test statistics is known or can be sampled from. When studied in a sparse detection setting, one of the proposed procedures reduces to a version of Tukey's "higher criticism" studied by Donoho and Jin (Ann. Stat., 2004) and thus is asymptotically optimal for detecting sparse heterogeneous mixtures. Our theoretical statements are supported by numerical experiments.

Email: pierre.neuvial@math.univ-toulouse.fr

WPM1-1 T1

Biomarker Subgroup Testing, Misclassification, and Missing Data

Gene Pennello, Jingjing Ye
FDA,USA

The objective of precision medicine has been stated as treating the "right patient with the right drug at the right time". Many predictive biomarkers facilitate precision medicine by explaining a clinically significant amount of the variation in a treatment effect. The anticipation that the treatment will only be effective in a biomarker-defined subgroup means that many proposed procedures for testing treatment effect overall and in one or more biomarker-defined subgroups are unsatisfactory. The clinical objective is not to find the largest population in whom statistical significance of the treatment effect is retained, but to determine the population (if it exists) in whom the effect is homogeneous and clinically significant. In this talk, we'll discuss frequentist and Bayesian testing procedures that have been designed to address the clinical objective of predictive biomarkers. We'll also quantify how biomarker measurement error attenuates the difference in treatment effect between biomarker defined subgroups. We'll also show that missing biomarker results (e.g., specimens unavailable or unevaluable for biomarker testing) can be addressed with Bayesian selection models even when minimal assumptions on the missing data mechanism mean that model parameters aren't fully identified.

Email: gene.pennello@fda.hhs.gov

WPM1-1 T2

Partitioning to Guarantee Subgroup Sensitive Inference in Personalized Medicine

Szu-Yu Tang
Ventana Medical Systems, Inc., USA

Many modern medicines are targeted therapies, targeting specific pathways. A binary biomarker associated with the pathway classifies patients into marker-positive (g+) and marker-negative (g-) subgroups. To decide whether to target the overall population ({g+, g-}), or only the marker-positive patients, or neither, inference on efficacy of the drug on the g+ patients, on the g-patients, and on the mixture {g+, g-} patients need to be assessed. There are logical relationships among efficacy parameters in g+, g-, and {g+, g-}. More important than any "power" consideration is that statistical inference should be Subgroup Sensitive in the sense that if efficacy in g+ and g- are inferred, then efficacy in g+, g- should automatically be inferred. (Subgroup Sensitivity implies Simpson's paradox does not occur.) This presentation first reviews the Subgroup Mixable concept, showing that Subgroup Sensitivity cannot be achieved if efficacy measure is Hazard Ratio or Odds Ratio. We then show, for Subgroup Mixable efficacy measures such as difference of means or ratio of medians, the Partition Principle in multiple testing recognizes Subgroup Sensitivity and thus would not form any intersection hypothesis contradicting it, thereby automatically guaranteeing Subgroup Sensitivity.

Email: tang.142@buckeyemail.osu.edu

WPM1-1 T3

Subgroup Finding via Bayesian Additive Regression Trees.

Siva Sivaganesan
University of Cincinnati, USA

We provide a Bayesian decision theoretic approach to finding subgroups that have elevated treatment effects. Our approach separates the modeling of the response variable from the task of subgroup finding, and allows a flexible modeling of the response variable irrespective of potential subgroups of interest. We use Bayesian additive regression trees (BART) to model the response variable, and use a utility function defined in terms of a candidate subgroup and the predicted response for that subgroup. Subgroups are identified by maximizing the expected utility where the expectation is taken with respect to the posterior predictive distribution of the response, and the maximization is done over an a priori specified set of candidate subgroups. Our approach allows subgroups based on both quantitative and categorical covariates. We illustrate the approach using simulated data sets and a real data set.

Email: sivagas@ucmail.uc.edu

WPM1-1 T4

Exploration of Heterogeneous Teatment Effects via Concave Fusion

Shujie Ma
University of California, Riverside, USA

Understanding treatment heterogeneity is essential to the development of precision medicine, which seeks to tailor medical treatments to subgroups of patients with similar characteristics. One of the challenges to achieve this goal is that we usually do not have a priori knowledge of the grouping information of patients with respect to treatment. To address this problem, we consider a heterogeneous regression model by assuming that the coefficients for treatment variables are subject-dependent and belong to different subgroups with unknown grouping information. We develop a concave fusion penalized method and derive an alternating direction method of multipliers algorithm for its implementation. The method is able to automatically estimate the grouping structure and the subgroup-specific treatment effects. We show that under suitable conditions the oracle least squares estimator with a priori knowledge of the true grouping information is a local minimizer of the objective function with high probability. This provides a theoretical justification for the statistical inference about the subgroup structure and treatment effects. We evaluate the performance of the proposed method by simulation studies and illustrate its application by analyzing the data from the AIDS Clinical Trials Group Study.

Email: shujie.ma@ucr.edu

WPM1-2 T1

Model-free Knockoffs for High-dimensional Controlled Variable Selection

Yingying Fan
University of Southern California, USA

A common problem in modern statistical applications is to select, from a large set of candidates, a subset of variables which are important for determining an outcome of interest. For instance, the outcome may be disease status and the variables may be hundreds of thousands of single nucleotide polymorphisms on the genome. For data coming from low-dimensional n> p linear homoscedastic models, the knockoff procedure recently introduced by Barber and Candes solves the problem by performing variable selection while controlling the false discovery rate (FDR). The present paper extends the knockoff framework to arbitrary (and unknown) conditional models and any dimensions, including n<p, allowing it to solve a much broader array of problems. This extension requires the design matrix be random (independent and identically distributed rows) with a covariate distribution that is known, although we show our procedure to be robust to unknown/estimated distributions. To our knowledge, no other procedure solves the variable selection problem in such generality, but in the restricted settings where competitors exist, we demonstrate the superior power of knockoffs through simulations. Finally, we apply our procedure to data from a case-control study of Crohn's disease in the United Kingdom, making twice as many discoveries as the original analysis of the same data.

Email: fanyingy@marshall.usc.edu

WPM1-2 T2

Multilayer False Discovery Rate Control for Variable Selection

Eugene Katsevich, Chiara Sabatti
Stanford University, USA

In certain applications, it is of interest to test the same set of hypotheses at different levels of granularity. Consider the setting of genome-wide association studies based on exome sequencing data, in which we seek genetic variants associated with a given trait. In exome sequencing data, each genetic variant belongs to a gene, so it is of interest both to discover a set of genetic variants (i.e. individual hypotheses) associated with the trait and a set of genes (i.e. groups of hypotheses) associated with the trait. Consider applying an FDR-controlling procedure at the level of genetic variants, and then reporting the list of discovered variants along with the genes to which they belong. The corresponding list of genes is of biological interest itself, but unfortunately comes with no FDR guarantees. In fact, in some cases the group false discovery rate can greatly exceed the individual false discovery rate. To remedy this problem, Barber and Ramdas (2016) propose a criterion called multilayer FDR control, a property of a selection procedure guaranteeing control at prespecified levels for more than one "layer." Barber and Ramdas propose the p-filter, a procedure obeying multilayer FDR control, given PRDS p-values for the individual-level hypotheses. In this talk, I will present the multilayer knockoff filter, a methodology that extends the p-filter concept to the setting in which the hypotheses are predictors in a (potentially high-dimensional) regression. The methodology is based on the framework of the knockoff filter (Barber and Candes 2015). Remarkably, the multilayer knockoff filter can actually gain power with respect to the regular knockoff filter in cases when the groups are informative with respect to the signal (i.e. there are

multiple non-null hypotheses per non-null group). Even when the groups are not informative, the multilayer knockoff procedure has similar power to the regular knockoff filter, while controlling both the individual and the group false discovery rates.

Email: katsevich.gene@gmail.com

WPM1-2 T3

Penalized Likelihood and Multiple Testing

Harold Sackrowitz, Arthur Cohen, John Kolassa
Rutgers University, USA

Multiple testing problems can be characterized by beginning with a collection of parameters. Then one individually tests each parameter to decide whether or not it is zero. The main focus of the literature is on the performance of the collection of these testing procedures. Yet, in many practical situations it would be of interest to follow the testing process by further inference on those parameters deemed different from zero. In variable selection problems one also has to decide which of a number of parameters are non-zero. Here the driving force is usually prediction based on estimates of the parameters. Little attention is given to the performance of the testing procedures themselves. In this talk we recognize the similarities in these two types of problem and try to exploit them. In particular, we consider the penalized likelihood methods that are very commonly used for model selection. We discuss how they would perform if used as multiple testing procedures in common multiple testing settings such as treatments versus control models.

Email: sackrowi@rci.rutgers.edu

WPM1-2 T4

Assessing Variable Selection Uncertainty in Linear Models

Aldo Solari, Ningning Xu, Jelle Goeman
University of Milano-Bicocca, Italy

The problem of variable selection in regression is old but still very relevant, and some recent progress has been made in this area. Notably, selective inference has been used to design new variable selection methods. Both old and new variable selection methods, however, tend to come up with very different models, especially in the presence of collinearity. This suggests that the uncertainty in the results of variable selection should be taken into account. In this talk we aim at quantifying the uncertainty in the variable selection process for linear models. Using the closed testing procedure, we construct a confidence set of models that covers (the best approximation of) the true model with (1-alpha) confidence, allowing for first-order model misspecification. We argue 1.) that such a confidence set represents the uncertainty in the variable selection process, and should always be taken into account when interpreting the results of a variable selection method; and 2.) that every admissible variable selection method should select a model from such a confidence set. The confidence set is characterized by its minimal elements, the minimal adequate models (MAMs). Usually the confidence set is spanned by a small number of MAMs, so that it is relatively easy to work with. We show that the proposed simultaneous inference approach is considerably less conservative than Scheffe protection. We focus on the definition of the null hypothesis of model adequateness and provide relationships with both old (Mallows 1973, Spjotvoll 1977,

etc.) and new (Berk et al. 2013, G'Sell et al. 2016, etc.) literature. Finally, we illustrate with classical examples how to construct the confidence set by using the cherry R package.

Email: aldo.solari@unimib.it

WPM1-3 T1

A Gatekeeping Procedure to Test a Primary and a Secondary Endpoint in a Group Sequential Design with Multiple Interim Looks

Ajit Tamhane, Jiangtao Gou, Christopher Jennison, Cyrus Mehta, Teresa Curto
Northwestern University, USA

Glimm et al. (2010) and Tamhane et al. (2010) studied the problem of testing a primary and a secondary endpoint, subject to gatekeeping constraint, using a group sequential design (GSD) with K = 2 looks. In this paper we greatly extend the previous results to multiple (K > 2) looks. The familywise error rate (FWER) is to be controlled at a pre-assigned level alpha. Obviously, the primary boundary must be alpha-level. We show under what conditions one alpha-level boundary is uniformly more powerful than another alpha-level boundary. Based on this result we recommend the choice of the OBrien-Fleming (1979) boundary over the Pocock (1977) boundary for the primary endpoint. For the secondary endpoint the choice of the boundary is more complicated since under certain conditions the secondary boundary can be chosen to have nominal level alpha' > alpha, thus allowing an increase in the secondary power. We carry out power comparisons via simulation between different choices of primary-secondary boundary combinations. The methodology developed in the paper is applied to the results from the RALES study (Pitt et al. (1999),Wittes et al. (2001)).

Email: atamhane@northwestern.edu

WPM1-3 T2

How to Evaluate Type II Error Rate with Multiple Endpoints

Bushi Wang, Naitee Ting
Boehringer Ingelheim, USA

The FDA draft guidance on multiple endpoints in clinical trials (January 2017) pointed out the regulatory concern of the type II error rate inflation with multiple endpoints. Many of the statistical adjustment to control the type I error rate for multiplicity decrease the study power because they lowered the alpha level used for each of the individual endpoints' test of hypothesis. The use of co-primary endpoints does not require multiplicity adjustment for type I error but will also increase the type II error rate and decrease study power. In this presentation, I provide detailed steps on how to evaluate sample size based on the objective of the clinical study and the selected multiplicity adjustment to control type I error. Analytic forms of power for individual endpoint hypothesis can be derived for most commonly seen scenarios. Simulation can be also easily set up. Optimal sample size is possible by fine tune the individual power for each endpoint with different effect size assumptions.

Email: bushi.wang@boehringer-ingelheim.com

WPM1-3 T3

Improved Testing Procedures for Group Sequential Trials with a Primary and a Secondary Endpoint

Huiling Li, Jianming Wang, Xiaolong Luo
Celegene, USA

In group sequential trials with a primary and a secondary endpoint, the Type I error rate for the primary endpoint is often controlled by choosing an alpha-spending function, e.g., the O'Brien-Fleming alpha-spending function. Given the selected alpha-spending function for the primary endpoint, we study an improved Bonferroni testing procedure and an improved Pocock testing procedure for the secondary endpoint. The improved procedures take into consideration the correlation between the interim and final statistics while applying graphical approaches and recycling significance levels from rejected hypothesis to an un-rejected one. Therefore, the resulting procedures improve the study power. The procedures control the family-wise error rate (FWER) in the strong sense by construction and then confirmed by simulations. We also compare the procedures with other commonly used existing procedures for the secondary endpoint, such as Bonferroni method and Pocock procedure. An example is provided to illustrate the properties of the procedures.

Email: huili@celgene.com

WPM1-3 T4

On Simultaneous Tests of Superiority and Noninferiority of Multiple Endpoints in Clinical Trials

Jie Chen,Tze L. Lai
Merck Research Laboratories, USA

Simultaneous tests of superiority and noninferiority of multiple endpoints in clinical trials are often conducted to demonstrate that a new treatment is superior on at least one endpoint and noninferior on the rest of the endpoints over a control. Several methods have been developed in the past decade or so to handle this type of superiority-noninferiority tests. This research extended Tamhane and Logan's work (2004) by using the Bonferroni inequality of Worsley (1982) to improve the critical boundary values that control the type I error probability at a desired level. Simulations are performed to compare the proposed approach with existing methods with respect to type I error rate control as well as study power. A real example is given to illustrate the application of the proposed approach.

Email: jie_chen@merck.com

WPM1-3 T5

Testing Superiority When Noninferiority of the Same Endpoint is Assessed in a Multiple Comparison Procedure

Scott Beattie, Jiajun Liu, Pedro Lopez-Romero
Eli Lilly and Company, USA

Background: It is common practice to assess non-inferiority (NI) of an experimental agent to an active standard-of-care medicine by an established endpoint and pre-specified NI margin during the conduct of a clinical trial with an active comparator. If such an assessment is a key feature of the trial, the hypothesis is included in the study's pre-specified multiple comparison procedure (MCP), along with other important hypotheses, to control the family-wise error rate (FWER). In the course of evaluating the evidence for NI, the data may indicate superiority in addition to NI. However, whether superiority may be claimed without inflation of the type I error rate if the related hypothesis for evaluating superiority had not been included in the MCP is a subject of frequent informal debate.
Objectives**:** To determine the conditions, if any, under which the FWER remains controlled in assessing the superiority null hypothesis for an endpoint when only the NI null hypothesis for that endpoint was included and rejected in a valid MCP involving other endpoints.
Methods and Results: The partitioning principle is used to evaluate several 1- and 2-endpoint scenarios in which the superiority hypothesis of interest is tested but only the related NI hypothesis is included within the pre-specified MCP. Some scenarios result in inflation of the FWER, whereas in others the error rate remains strongly controlled. Thus in general the superiority null hypothesis cannot be tested without inflating the type I error rate if it was not included within the MCP. However, the closed testing principle is used to prove that the FWER is preserved and superiority can be concluded within the bounds of strong control under the following conditions:
_ All null hypotheses included in the MCP are rejected,
_ There is at most 1 such related superiority hypothesis and it is not considered to be arbitrarily chosen.
Conclusio*n*: Rejection of the non-inferiority null hypothesis within a multiple comparison procedure may lead naturally to an assessment of the superiority null hypothesis for that endpoint. A conclusion of superiority is statistically valid when all null hypotheses in the MCP are rejected and the superiority hypothesis related to a rejected non-inferiority null hypothesis is unique.

Email: scottbt@lilly.com

WPM2-1 T1

Online Rules for Control of False Discovery Rate

Adel Javanmard
University of Southern California, USA

Multiple hypothesis testing is a core problem in statistical inference and arises in almost every scientific field. For a given set of null hypotheses, Benjamini and Hochberg introduced the notion of false discovery rate (FDR), which is the expected proportion of false positives among rejected null hypotheses, and further proposed a testing procedure that controls FDR below a pre-assigned significance level. Nowadays FDR is the criterion of choice for large-scale multiple hypothesis testing. In this talk, we consider the problem of controlling FDR in an "online manner". Concretely, we consider an ordered, possibly infinite, sequence of null hypotheses where at each step the statistician must decide whether to reject current null hypothesis having access only to the previous decisions. We introduce a class of generalized alpha-investing procedures and prove that any rule in this class controls FDR in online manner.

Email: ajavanma@usc.edu

WPM2-1 T2

Sequential Testing of Multiple Hypotheses Under Arbitrary Joint Distributions

Michael Hankin, Jay Bartroff
University of Southern California, USA

We develop procedures for sequential testing of multiple hypotheses with False Discovery Rate control under arbitrary dependence conditions. Our work corrects and extends Rao and Guo's optimization-based proofs of fixed sample size FDR control of step-down procedures into the sequential case, as initially developed by Bartroff and Song. To do so we decompose FDR into a sum of the probabilities of a carefully chosen set of random events. The random events depend on the sequential rejection procedure, and the probabilities depend on the joint distribution of test statistics. We then maximize FDR over all possible joint distributions, subject to the usual marginal conditions employed in sequential testing procedures, and use that value to scale the maximum FDR to our desired value. We further extend these results to FDR and FNR control under infinite horizon procedures as well as finite and infinite horizon pFDR (and pFNR) controlling procedures. After proving the validity of our claims analytically, we simulate our procedure using the UK's Yellowcard Drug Side Effect Report database to detect drugs that may cause amnesia as well as running it on purely synthetic data where the ground truth is known which allows for careful analysis of its operating characteristics.

Email: meh2135@gmail.com

WPM2-1 T3

Methods for Multiple Testing Error control on Sequential Data

Jay Bartroff
University of Southern California, USA

I will review recent work on procedures for multiple testing on sequential data, which is the natural setting of many applications like multi-endpoint clinical trials, high-throughput gene sequencing technologies, biosurveillance, sequential cross validation in high dimensional models, and pharmacovigilance reporting systems. I will review sequential multiple testing procedures for FWER and FDR control, as well as popular generalizations like their tail probabilities and pFDR. The sequential procedures are flexible in that they can be applied to data streams of arbitrary dimension and dependence and arbitrary null hypotheses, requiring only that the conventional type I and II error probabilities can be controlled marginally. I will give tips for implementation and discuss some real-data applications.

Email: bartroff@usc.edu

WPM2-1 T4

Sequential Multiple Testing with Generalized Error Control: An Asymptotic Optimality Theory

Yanglei Song, Georgios Fellouris
University of Illinois at Urbana-Champaign, USA

The multiple testing problem is considered under two different error metrics, when the data for the various hypotheses are collected sequentially in independent streams. In the first one, the probability of making at least k mistakes, of any kind, is controlled. In the second, the probabilities of at least k1 false positives and at least k2 false negatives are simultaneously controlled below two arbitrary levels. For each formulation, we characterize the optimal expected sample size to a first-order asymptotic approximation as the error probabilities vanish (at arbitrary rates). More importantly, for each formulation we propose a novel, feasible sequential multiple testing procedure that achieves the optimal asymptotic performance under every possible signal configuration. These asymptotic optimality results are established under weak distributional assumptions which hold beyond the case of i.i.d. observations in the streams.

Email: ysong44@illinois.edu

WPM2-2 T1

Statistical Pattern Mining: An Overview

Koji Tsuda
University of Tokyo, Japan

To discover new knowledge from a large amount of data, pattern mining techniques such as item set mining, sequence mining and graph mining have been applied to a wide range of problems. To convince biomedical researchers, however, it is necessary to show statistical significance of obtained patterns to prove that the patterns are not likely to emerge from random data. The key concept of significance testing is family-wise error rate, i.e., the probability of at least one pattern is falsely discovered under null hypotheses. In the worst case, FWER grows linearly to the number of all possible patterns. We show that, in reality, FWER grows much slower than the worst case, and it is possible to find significant patterns in biomedical data. The following two properties are exploited to accurately bound FWER and compute small p-value correction factors. 1) Only closed patterns need to be counted. 2) Patterns of low frequency can be ignored, where the frequency threshold depends on Tarone's minimum achievable significance level. In this talk, I review the emerging field of statistical pattern mining, highlighting new algorithmic techniques to apply multiplicity control to combinatorially many hypotheses. The techniques allow us to apply conventional statistical tests to unconventional data types including sets, sequences and graphs that are prevalent in modern natural and social sciences.

Email: tsuda@k.u-tokyo.ac.jp

WPM2-2 T2

Selective Inference for Predictive Pattern Mining

Ichiro Takeuchi, Shinya Suzumura, Yuta Umezu, Koji Tsuda
Nagoya Institute of Technology, Japan

Discovering statistically significant patterns from databases is an important challenging problem.The main obstacle of this problem is in the difficulty of taking into account the selection bias, i.e., the bias arising from the fact that patterns are selected from extremely large number of candidates in databases. In this paper, we introduce a new approach for predictive pattern mining problems that can address the selection bias issue. Our approach is built on a recently popularized statistical inference framework called selective inference. In selective inference, statistical inferences (such as statistical hypothesis testing) are conducted based on sampling distributions conditional on a selection event. If the selection event is characterized in a tractable way, statistical inferences can be made without minding selection bias issue. However, in pattern mining problems, it is difficult to characterize the entire selection process of mining algorithms. Our main contribution in this talk is to solve this challenging problem for a class of predictive pattern mining problems by introducing a novel algorithmic framework. We demonstrate that our approach is useful for finding statistically significant patterns from databases.

Email: takeuchi.ichiro@nitech.ac.jp

WPM2-2 T3

Accounting for a Categorical Covariate in Significant Pattern Mining

Llinares Lopez Felipe, Laetitia Papaxanthos, Dean Bodenham, Damian Roqueiro, Karsten B
ETH Zurich, Switzerland

Recent work has shown that, by combining Tarone's improved Bonferroni correction for discrete data with the apriori property of pattern mining, it is possible to efficiently explore all combinations of features while guaranteeing FWER control. Despite their success, a major pitfall of this family of pattern mining algorithms has been their inability to account for covariates, limiting their applicability in domains such as computational biology or healthcare. In this talk, a novel approach that allows correcting for a categorical covariate will be presented. Empirically, the resulting algorithm drastically reduces the number of false positives found due to confounding without sacrificing statistical power or computational efficiency. An application of this method to aggregate weak effects within arbitrary genomic regions in genome-wide association studies will be discussed. Unlike existing approaches, such as burden tests, the resulting method does not require prior specification of a small set of genomic regions to be tested. Rather, it is able to test all genomic regions, regardless of size or starting position. This leads to increased statistical power in settings where prior knowledge is not available to confidently narrow down the set of genomic regions of interest or to estimate their optimal size.

Email: felipe.llinares@bsse.ethz.ch

WPM2-2 T4

Controlling Familywise Error When Rejecting at Most One Null Hypothesis Each From a Sequence of Sub-Families of Null Hypotheses

Geoff Webb, Mark van der Laan
Monash University, Australia

We present a procedure for controlling FWER when sequentially considering successive subfamilies of null hypotheses and rejecting at most one from each subfamily. This scenario arises in stepwise model selection. Our procedure differs from previous procedures for controlling FWER by adjusting the critical values that are applied in subsequent rejection decisions by subtracting from the global significance level alpha quantities based on the p-values of rejected null hypotheses and the numbers of null hypotheses considered.

Email: geoff.webb@monash.edu

WPM2-3 T1

Application of Frequentist Guidelines in Bayesian Adaptive Designs

Jian Zhu,Yi Liu
Takeda Pharmaceuticals, USA

Bayesian adaptive designs are becoming popular in clinical trials, which intend to gain efficiency and ethical advantages by adopting stopping rules and/or adaptive randomization in multiple interim analyses. However, unlike frequentist designs such as group sequential designs, the thresholds for interim decision-making in Bayesian adaptive designs are usually determined by simulations: for each design, statisticians try different combination of values until they find the threshold set that yields the desirable operating characteristics. To improve this procedure, we apply various frequentist methods such as alpha spending function and multiple comparison procedures for such designs. We study an example with continuous endpoints and construct two equivalent versions within Bayesian and frequentist framework respectively. Through this example we demonstrate that, 1. It is more efficient to determine the thresholds in Bayesian designs by frequentist methods; 2. The chosen thresholds are more intuitive and easier to understand; 3. The hybrid designs are more comparable.

Email: jian.zhu2@takeda.com

WPM2-3 T2

Blinded Sample Size Re-Estimation in Three-Arm Trials with 'Gold Standard' Design

Tobias Mütze, Tim Friede
University Medical Center Gottingen, Germany

The sample size of a clinical trial relies on information about nuisance parameters such as the outcome variance. When no or only limited information is available, it has been proposed to include an internal pilot study in the design of the trial [1]. Based on the results of the internal pilot study, the initially planned sample size can be adjusted. In this contribution we present results of our study of blinded sample size re-estimation in the 'gold standard' design with internal pilot study for normally distributed outcomes. The 'gold standard' design is a three-arm clinical trial design which includes an active and a placebo control in addition to an experimental treatment [2,3]. We focus on the absolute margin approach to three-arm trials at which the performance of the experimental treatment and the assay sensitivity are assessed by pairwise comparisons [4]. We compare several sample size re-estimation procedures, in particular the procedure based on the one-sample approach which is recommended for two-arm trials [5] and the procedure based on the Xing-Ganju approach [6], in a simulation study assessing operating characteristics including power and type I error rate. The simulation study shows that sample size re-estimation based on the popular one-sample variance estimator results in overpowered trials. Moreover, sample size re-estimation based on unbiased variance estimators such as the Xing-Ganju variance estimator results in underpowered trials, as it is expected since an overestimation of the variance and thus the sample size is in general required for the re-estimation procedure to eventually meet the target power. To overcome this problem, we propose an inflation factor for the sample size re-estimation with the Xing-Ganju variance estimator and show that this approach results in adequately powered trials. Due to favorable features of the Xing-Ganju variance estimator such as unbiasedness and a distribution independent of the group means, the inflation factor does not depend on the nuisance parameter and, therefore, can be calculated prior to a trial. Moreover, we prove that the proposed sample size re-estimation procedure based on the Xing-Ganju variance estimator does not bias the effect estimator at the end of the trial, in contrast to the sample size re-estimation based on the one-sample variance estimator.

References

[1] Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. Statistics in Medicine, 9:65-72.
[2] Pigeot, I., Schafer, J., Rohmel, J., and Hauschke, D. (2003). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. Statistics in Medicine, 22: 883-899.
[3] Koch, A. and Rohmel, J. (2004). Hypothesis testing in the "gold standard" design for proving the e_cacy of an experimental treatment relative to placebo and a reference. Journal of Biopharmaceutical Statistics, 14:315-325.
[4] Stucke K. and Kieser M. (2012). A general approach for sample size calculation for the three-arm 'gold standard' non-inferiority design. Statistics in Medicine 31:3579-3596.
[5] Friede T. and Kieser M. (2013). Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation. Pharmaceutical Statistics. 12:141-146.
[6] Xing B. and Ganju J. (2005). A method to estimate the variance of an endpoint from an 48 on-going blinded trial. Statistics in Medicine 24:1807-1814.

Email: tobias.muetze@med.uni-goettingen.de

WPM2-3 T3

Optimized Adaptive Enrichment Designs for Clinical Trials with a Sensitive Subpopulation

Martin Posch, Thomas Ondra, Sebastian Jobjoernsson, Carl-Fredrik Burman, Franz Koenig, Nigel S
Medical University of Vienna, Austria

It has been proposed to optimize single stage clinical trials with a sensitive subpopulation by optimizing a utility function weighting gains and costs of a particular trial design. We extend the current literature in several ways. First, we consider partially enriched designs, where the subgroup prevalence in the trial may differ from the subgroup in the underlying population. Furthermore we derive, for utility functions representing a sponsor's and a societal perspective, optimal adaptive two stage enrichment trials. Via a dynamic programming approach we optimize the first stage sample size and subgroup prevalence and derive optimal data dependent adaptation rules to select the second stage population, sample size and subpopulation prevalence. We show that in many of the investigated scenarios adaptive enrichment designs lead to a higher expected utility than single stage designs for both the sponsor and the societal perspective. Furthermore we demonstrate that adaptive enrichment designs are less sensitive with respect to the choice of the prior distribution as compared to single stage designs.

Email: martin.posch@meduniwien.ac.at

WPM2-3 T4

Optimal Adaptive Enrichment Trials

Thomas Burnett
University of Bath, UK

When conducting confirmatory clinical trials it may be possible to identify sub-populations of patients who respond differently to the new treatment. Adaptive Enrichment trials aim to make efficient use of pre-identified patient sub-populations; initially patients are sampled from all sub-populations, then at an interim analysis sub-populations are selected for the remaining sample. We test the null hypotheses corresponding to the remaining sub-populations after the interim analysis. To ensure strong control of the FamilyWise Error Rate (FWER) for all combinations of null hypotheses, we make use of closed testing procedures and combination tests. Since strong control of the FWER is ensured for all possible selections of patient sub-populations, at the interim analysis we are able to select these sub-populations by any method we choose. We use a Bayesian decision framework to optimize this decision at the interim analysis. We define a prior distribution accounting for uncertainty about the true treatment effects and a gain function to give a single measure of trial performance. At the interim analysis we select the sub-populations that yield the highest expected gain for the remainder of the trial. Using the Bayes expected gain we find the overall performance of the Bayes optimal Adaptive Enrichment designs, comparing this with alternative fixed sampling designs to learn about the benefits of the adaptive design. This structure for optimizing Adaptive Enrichment trials is easily applied to delayed response, longitudinal data and survival endpoints where information on short term observations may be used to enhance the decision with no impact on the FWER.

Email: t.burnett@bath.ac.uk

WPM2-3 T5

Multi-A(rmed)/B(andit) testing with Online FDR Control

Fanny Yang, Aaditya Ramdas, Kevin Jamieson, Martin Wainwright
University of California, Berkeley, USA

We propose a new framework as an alternative to existing setups for controlling false alarms across multiple A/B tests. It combines ideas from online false discovery rate (FDR) control with pure exploration for best-arm identification in multi-armed bandits (MAB). This framework has various applications, including pharmaceutical companies testing a control pill against a few treatment options, to internet companies testing their current default webpage (control) versus many alternatives (treatment). Our setup involves running a possibly infinite sequence of best-arm MAB instances, and controlling the overall FDR of the process in a fully online manner. Our main contributions are: (i) to propose reasonable definitions for a null hypothesis; (ii) to demonstrate how one can derive an always-valid sequential p-value for such a null hypothesis which allows users to continuously monitor and stop any running MAB instance at any time; and (iii) to embed MAB instances within online FDR algorithms in a way that allows setting MAB confidence-levels based on FDR rejection thresholds. In addition, we adapt existing theory from both the MAB and online FDR literature to ensure that our framework comes with strong sample-optimality guarantees, as well as control of the power and (a modified) FDR at any time. We run extensive simulations to verify our claims, and also report results on real data collected from the New Yorker Cartoon Caption contest.

Email: fanny-yang@berkeley.edu

WPM3-1 P1

Statistical Properties of Bernstein copulae with Applications in Multiple Testing

Andre Neumann, Taras Bodnar, Dietmar Pfeifer, Thorsten Dickhaus
University of Bremen, Germany

A general way to estimate continuous functions consists of approximations by means of Bernstein polynomials. Sancetta and Satchell (2004) proposed to apply this technique to the problem of approximating copula functions. The resulting so-called Bernstein copulae are nonparametric copula estimates with some desirable mathematical features like smoothness. We extend previous statistical properties regarding bivariate Bernstein copulae for the multivariate case and present their impact on multiple tests. In particular, we utilize them to derive asymptotic confidence regions for the family-wise error rate (FWER) of simultaneous test procedures which are empirically calibrated by making use of Bernstein copulae approximations of the dependency structure among the test statistics. This extends a similar approach by Stange et al. (2015) in the parametric case. A simulation study quantifies the gain in FWER level exhaustion and, consequently, power which can be achieved by exploiting the dependencies, in comparison with common threshold calibrations like the Bonferroni or the Sidak correction.

Email: neumann@uni-bremen.de

WPM3-1 P2

Simultaneous Confidence Intervals for Pairwise Comparisons Among Mean Vectors With Monotone Missing Data

Ayaka Yagi, Takashi Seo
Tokyo University of Science, Japan

Simultaneous confidence intervals for pairwise multiple comparisons among mean vectors when each dataset has a monotone missing data pattern are considered. The maximum likelihood estimators (MLEs) of the mean vector and the covariance matrix in the case of monotone missing data under multi-sample problem are derived. An approximate upper percentile of the simplified Hotelling's T_2 statistic, and that of the T_{max}^2 type statistic by Bonferroni's approximation procedure are presented in the case of monotone missing data. Approximate simultaneous confidence intervals for pairwise comparisons among mean vectors are also presented under two-sample and multi-sample problems. Finally, the accuracy and asymptotic behavior of the approximation are investigated by Monte Carlo simulation.

Email: 1415701@ed.tus.ac.jp

WPM3-1 P3

Constructing Tests to Compare Two Proportions Whose Critical Regions Guarantee to be Barnard Convex Sets

Jose Juan Castro Alva, Felix Almendra-Arao, Hortensia Josefina Reyes-Cervantes
Facultad de Ciencias Fisico Matematicas de la Benemerita Universidad Autonoma de, Mexico

In the context of non-inferiority (NI) tests and superiority (S) tests, the critical region must be a Barnard convex set (BCS) because of two main reasons. One is about the compute of test size, based on the fact that calculating test size is a computational intensive problem, however the computational time which is involved in computing the test size its reduce when the critical region is a Barnard convex set. The second reason is that in order for the NI/S statistical tests to make sense. Due to it is indeed possible for NI/S tests' critical regions to not be Barnard, it is desirable that they are for the reasons stated above. Therefore, it is important to generate, from a given NI/S test, a test, which guarantees that the critical regions are Barnard convex sets. We propose a method by which, from a given NI/S test, we construct another NI/S test, ensuring that the critical regions corresponding to the modified test are Barnard convex sets, we illustrate this through examples.

Email: jjcasatroa@gmail.com

WPM3-1 P4

A Frequency-Domain Model Selection Criterion for a (Dynamic) Factor Model

Natalia Sirotko-Sibirskaya
University of Bremen, Germany

We consider a (dynamic) factor model such that a signal of dimension 'p' can be decomposed into common components of dimension 'k' and idiosyncratic components of dimension 'p'. This decomposition is used when it is assumed that 'k' is much smaller than 'p', therefore, the key step is in finding the rank 'k' of such lower rank representation, so that the dynamics of the original process can be recovered by such lower-rank representation. Up to now there is no unanimous agreement among the researchers on which method to use in order to choose the optimal number of factors. Classical methods include computing likelihood-ratio tests and using screenplots in principal-component analysis, however, these methods impose an assumption of homoscedastic noise of idiosyncratic factors which can be regarded as limiting in view of the current research. Recent developments include AIC/BIC type of criteria adaptation to factor model analysis, see Bai and Ng (2002), dynamic principal component analysis, see Hallin and Liska (2007), bi-cross-validation, see Owen and Wang (2005), and others. We propose a data-driven method for selecting an "optimal" number of factors. The method is based on cross-validation technique for a factor model evaluated in the frequency domain and allows to relax the assumption of homoscedasticity of idiosyncratic components. In the spirit of Hurvich and Zeger (1990) we define a frequency-domain-cross validation criterion, in this case, for a factor model. It can be shown that expectation of a frequency-domain cross-validation criterion is approximately equal to the sum of the MSE of a spectrum estimate and variance of idiosyncratic components. This criterion is evaluated for each possible choice of k. The choice of the "optimal" model is based on minimization of the corresponding criterion. The proposed method is then compared to several existing criteria in Monte-Carlo simulations as well as applied to a data set to evaluate its performance.

Email: sirotkos@uni-bremen.de

WPM3-1 P5

A Computer-Assisted Pap Smear Screening System Based on Automated Cell Nuclei Segmentation

Fang-Hsuan Cheng, Nai-Ren Hsu
Chung Hua University, Taiwan

Malignant tumor, also known as carcinoma, is the first of top 10 causes of death in which the cervical carcinoma is the top 5 common cancer of women. With the popularization of Pap test, the rank of cervical carcinoma has a declining trend. The prevention of cervical carcinoma depends on early detection and treatment. Pap test is the most effective method for early screening. With the increase of the screening rate, there were more and more workload to the doctors and medical staff. Subjective view, heavy duty and overworked were causing the mistakes of the screening. Therefore, the automatic computer-aided system for Pap test is the new trend to solve it. In this paper, we used the Bethesda system, a system for reporting cervical or vaginal cytological diagnoses, as the basis of screening, and image processing and computer vision method are applied to retrieve the feature of abnormal cells. Carcinoma cell nucleus segmentation is the key step to automated screening system for Pap test of cervical smear. In this study, we segment the cell of smear image into nucleus and cytoplasm in HSV color space and calculate the global nuclear-cytoplasm ratio. Next, we find the contour of nuclei by morphological expansion and erosion methods. The deformation features are also recorded in this step. Finally, area features of the Syncytium-like cell and color characteristics of the Hyperchromasia cell are estimated. Combine all of the above features, we mark the tumor location with color circle and block as a reference for doctors and medical staff. The experimental results show that the accuracy of the proposed system is 0.975, sensitivity is 0.974 and the specificity is 1 of screening the images into normal and abnormal ones. Furthermore, the classification accuracy of normal cell is 1, LSIL cell is 0.7 and HSIL cell or Cancer is 0.72. It is

concluded that the proposed system can screen the tumor cell automatically and help the medical staffs to do their work.

Email: fhcheng@chu.edu.tw

WPM3-1 P6

Control of False Discoveries in Grouped Hypothesis Testing for eQTL Data

Pratyaydipta Rudra, Andrew Nobel, Fred A. Wright
University of Colorado Denver, Anschutz Medical Campus, USA

Expression quantitative trait loci (eQTL) analysis aims to detect the loci that influence the expression of one or more genes. The gene expression is considered as the quantitative trait potentially associated with the genotypes at different sites in the genome that are usually various single nucleotide polymorphisms (SNPs). We describe a statistical method for testing hypotheses that form into groups, with each group showing potentially different characteristics. Methods to control family-wise error rate or false discovery rate for group testing have been proposed earlier, but may not easily apply to expression quantitative trait loci (eQTL) data, for which certain structured alternatives may be defensible and enable the researcher to avoid overly conservative approaches. In an empirical Bayesian setting, we propose a new method to control the false discovery rate (FDR) for grouped hypothesis data. Here, each gene forms a group, with SNPs annotated to the gene corresponding to individual hypotheses. Heterogeneity of effect sizes in different groups is considered by the introduction of a random effects component. Our method, entitled Random Effects model and testing procedure for Group-level FDR control (REG-FDR) assumes a model for alternative hypotheses for the eQTL data and controls the FDR by adaptive thresholding. We also propose Z-REG-FDR, an approximate version of REG-FDR that uses only Z-statistics of association between genotype and expression for each gene-SNP pair. Simulations demonstrate that Z-REG-FDR performed similarly to REG-FDR, but with much improved computational speed.

Email: pratyaydipta.rudra@ucdenver.edu

WPM3-1 P7

FDR Control for Dependent Chi-Square Goodness of Fit Tests

Melinda McCann, Amy Wagler
Oklahoma State University, USA

Somerville (2004) developed step-up and step-down FDR procedures that are valid for both dependent and independent hypotheses. These procedures involve calculating critical values under a series of least favorable configurations. The Somerville (2004) procedure requires that the hypotheses of interest involve testing a location parameter, utilizes this parameter to obtain the appropriate series of least favorable configurations, and implements the procedure by assuming a multivariate normal distribution and simulating random variables to estimate appropriate critical values. We utilize a similar approach for controlling FDR in situations where we are testing i = 1,…, m hypotheses using chi-square goodness-of-

fit tests. For this situation, the ith least favorable configuration can be determined. Accordingly, the appropriate critical values can be estimated by simulating multivariate chi-square random variables to estimate the appropriate critical values. We investigate the performance of this method and compare it to the Benjamini-Hochberg (1995) approach utilizing chi-square critical values.

Email: mccann@okstate.edu

WPM3-1 P8

Multiplicity Correction in Group-Sequential Oncology Trials Including Subgroup Analyses and Multiple Primary Endpoints

Agnes Balogh
Bristol-Myers Squibb, USA

A hypothetical oncology clinical trial design will be presented with several sources of multiplicity: comparing multiple regimens to control arm, multiple primary endpoints (overall survival and progression-free survival), interim analyses, subgroup analyses. In the framework of multiple testing in group-sequential trials using graphical approaches (Maurer and Bretz 2013), extensive simulation study has been performed to compare the different design and testing strategy options. The main purpose is not to miss opportunities for early success based on a surrogate endpoint or based on a larger efficacy signal of the definite endpoint, while keeping the probability of the final success of the trial.
Reference
1. Willi MAURER and Frank BRETZ: Multiple Testing in Group Sequential Trials Using Graphical Approaches. American Statistical Association Statistics in Biopharmaceutical Research November 2013, Vol.5, No.4)

Email: agnes.balogh@bms.com

WPM3-1 P9

Online FDR Control with Decaying Memory and Weights

Fanny Yang, Aaditya Ramdas, Martin Wainwright, Michael Jordan
University of California, Berkeley, USA

In the online multiple testing problem, p-values corresponding to different null hypotheses are presented one at a time, and the decision of whether to reject or not must be made immediately. This setup was proposed by Foster and Stine, and their alpha-investing algorithms have since been applied to various settings, like quality-preserving databases for science and multiple A/B tests for internet commerce. However, when used for large periods of time, these algorithms may suffer from a “piggybacking” problem, where a string of (possibly true) discoveries at one time can possibly cause a string of (possibly false) discoveries at a much later time, an undesirable effect in some applications. Our main contribution is to improve the class of generalized alpha-investing algorithms (GAI) in four orthogonal ways — (a) we award larger alpha-wealth for rejections under independence, implying higher power while maintaining FDR control, (b) we incorporate weights to indicate prior knowledge of which hypotheses are likely to be null or non-null, (c) we allow for differing penalties for false discoveries to indicate that some hypothesis tests may be more important than others, (d) we introduce a discount factor that indicates a decaying memory for past decisions, and define the decaying memory false discovery rate, or memFDR that directly addresses the piggybacking problem. Our GAI++ algorithms incorporate (a, b, c , d) simultaneously, and reduce to more powerful variants of earlier algorithms when the weights and decay are all set to unity.

Email: fanny-yang@berkeley.edu

WPM3-1 P10

Comparison of Different Variable Selection Methods in a Special Situation

Ningning Xu
Leiden University Medical Center, Netherlands

A wide variety of methods have been brought to solve the problem of variable selection and model selection for statistical applications, such as stability selection method, closed testing procedure and also best subsets, forward stepwise, backward stepwise. However, if there exists a strong noise variable for the response variable, some of the methods may not function optimally for statistical inference. We consider a linear regression model with a response variable and p explanatory variables where there exist two true variables X_1, X_2 and a strong noise variable $X_3 = X_1 + X_2 + \varepsilon_3$, which is more highly correlated with the variable of interest than the true ones. Variable selection methods such as stability selection method always choose the noise variable rather than the real ones as the true variables, while the closed testing procedure, which does not select a single model but a range of potential models, shows its superiority in this case as true variables are included in one of the potential models. Therefore, we would like to show that closed testing procedure works better than the stability selection method in this situation. And simulation results for best subset method, forward stepwise and backward stepwise also will be seen in our work.

Email: xu15263142750@gmail.com

WPM3-1 P11

Decentralized Decision Making on Networks with False Discovery rate Control

Jianbo Chen, Aaditya Ramdas, Michael Jordan, Martin Wainwright
University of California, Berkeley, USA

The field of distributed computation, learning, testing and inference on graphs has witnessed large advances in theory and wide adoption in practice. However, there do not currently exist any methods for multiple hypothesis testing on graphs that are equipped with provable guarantees on error metrics like the false discovery rate (FDR). In this talk, we consider a novel but natural setting where distinct agents reside on the nodes of an undirected graph, and each agent possesses p-values corresponding to one or more hypotheses local to its node. Each agent must individually decide whether to reject one or more local hypotheses by only communicating with its neighbors, with the joint aim that the global FDR over the entire graph must be controlled at a predefined level. We propose a simple decentralized family of

Query-Test-Exchange (QuTE) message-passing algorithms that interpolate smoothly between the two extremes — with zero communication budget, QuTE reduces to a simply Bonferroni correction, and with an unbounded communication budget or a centralized server node, QuTE reduces to the classical Benjamini-Hochberg (BH) procedure. Our main theoretical result is that the overall FDR is controlled *at any time* during the dynamic communication process, and that the power increases monotonically with communication, achieving the power of the gold-standard BH procedure after a time equaling the graph diameter. We explain how to deal with quantization issues involved in communicating real p-values, by introducing a new concept called p-ranks. We study the power of our procedure using a simulation suite of different levels of connectivity and communication on a variety of graph structures, and also provide an illustrative real-world example on an indoor sensor network.

Email: jianbochen@berkeley.edu

WPM3-1 P12

Post Selection Inference with Kernels

Yuta Umezu, Makato Yamada, Kenji Fukumizu, Ichiro Takeuchi
RIKEN, Japan

We propose a novel kernel based post selection inference (PSI) algorithm, which can not only handle non-linearity in data but also structured output such as multi-dimensional and multi-label outputs. Specifically, we develop a PSI algorithm for independence measures, and propose the Hilbert-Schmidt Independence Criterion (HSIC) based PSI algorithm (hsicInf). The novelty of the proposed algorithm is that it can handle non-linearity and/or structured data through kernels. Namely, the proposed algorithm can be used for wider range of applications including nonlinear multi-class classification and multivariate regressions, while existing PSI algorithms cannot handle them. Through synthetic experiments, we show that the proposed approach can find a set of statistically significant features for both regression and classification problems. Moreover, we apply the hsicInf algorithm to a real-world data, and show that hsicInf can successfully identify important features.

Email: umezu.yuta@nitech.ac.jp

WPM3-1 P13

Interactive Accumulation Test: A flexible framework for structural multiple testing.

Lihua Lei, William Fithian
University of California, Berkeley

Abstract: We consider the problem of controlling FDR for structural multiple testing where the rejected hypotheses should satisfy certain combinatorial constraints. We propose a general framework, as a generalisation of Accumulation Test (Li and Barber 2015), that we refer to as Interactive Accumulation Test (IAT). IAT uses the idea of "data carving" in the selective inference literature to divide the information contained in p-values into two parts: one for learning the underlying structure and the other for testing, where the division rule is derived from solving an ordinary differential equation. Under the

standard assumptions, we prove that IAT controls FDR in finite samples. IAT is a highly flexible framework that is able to deal with most structural multiple testing problems in principle. In particular, we consider three problems in this paper: 1) detecting a convex region from the space where the hypotheses lie on; 2) hierarchical testing; 3) selecting under the strong/weak heredity principle on a directly acyclic graph. We confirm and complement our theory via extensive numerical studies.

Email: lihua.lei@berkeley.edu

ThAM1-1 T1

Keeping the Blind Blind

Janet Wittes
Statistics Collaborative, Inc., USA

On paper, blinded adaptation seems fine: one looks at blinded (aka, masked) data, compares it to assumptions, and adapts the sample size accordingly. Or, someone in a fire-walled room, looks at unblinded data, makes a recommendation for adaptation (but describes the reason in a way that maintains the blind for those outside the firewall). We statisticians know how to characterize the operating characteristics of these procedures so that we preserver the Type I error rate of the trial and maintain its power. In designing such trials, however, we must be thoughtful about the operational consequences of these adaptations. Examples illustrate the difficulties.

Email: janet@statcollab.com

ThAM1-1 T2

Blinded Adaptations of Clinical Trials - How Blind Do We Have to Be?

Ekkehard Glimm
Novartis Pharma AG, Basel, Switzerland

Planned or unplanned modifications of ongoing clinical trials are sometimes unavoidable. It is often claimed that such modifications are unproblematic as long as they are done in a "blinded" fashion or "prior to data base lock". These terms seem to suggest that the personnel dealing with running and analyzing the trial has no information whatsoever regarding the data obtained in the trial so far. In reality, however, the term "blinded sample size modification" merely indicates that the treatment assignment is hidden to the clinical trial team. Any information not depending on this, for example the total variance, the total number of observed events, the overall event rate or the total average response is then available for modifications of the trial "free of charge". In this talk, we will discuss some popular types of blinded modifications and critically assess under what assumptions it is justified to treat them as if they had been determined prior to the collection of any data.

Email: ekkehard.glimm@novartis.com

ThAM1-1 T3

Bayesian Sample Size Re-estimation Incorporating External Data

Tobias Mütze, Tim Friede
University Medical Center Gottingen, Germany

Current methods on sample size re-estimation, in particular blinded sample size re-estimation, only utilize data from the internal pilot study to adjust the sample size of an on-going trial. However, external data from related studies is often already available. We propose methods to incorporate external data into sample size re-estimation and discuss the merits of the proposed methods compared to the traditional approaches of sample size re-estimation.

Email: tobias.muetze@med.uni-goettingen.de

ThAM1-1 T4

Estimation Following Blinded Adaptation

Michael Proschan
NIAID, NIH, USA

A re-randomization test can be used to salvage results of a trial that undergoes an unplanned change. The p-value remains valid under a strong null hypothesis, but the next question is whether there are any valid estimation methods following an unplanned change. This talk shows that estimation in this context is fraught with difficulty because although the treatment assignment vector and data may be independent under a strong null hypothesis, they are not independent under an alternative hypothesis. Nonetheless, some estimation method must be used, so what is the least objectionable method? This talk will attempt to answer that question.

Email: proscham@niaid.nih.gov

ThAM1-2 T1

A Modified Benjamini-Hochberg Procedure for Discrete Data

Sebastian Doehler, Guillermo Durand, Etienne Roquain
Darmstadt Univerity of Applied Sciences, Germany

The Benjamini-Hochberg procedure is a classical method for controlling the false discovery rate for multiple testing problems. This procedure was originally designed for continuous test statistics. However, in many applications, such as the analysis of next-generation sequencing data, the test statistics are discretely distributed. While it is well known that the Benjamini-Hochberg procedure still controls the false discovery rate in the discrete paradigm, it may be unnecessarily conservative. In this talk we aim to improve the Benjamini-Hochberg procedure in such settings by incorporating the discreteness of the p-value distributions. We investigate the performance of these approaches for empirical and simulated data.

Email: sebastian.doehler@h-da.de

ThAM1-2 T2

Discrete FDR Method Increases Sensitivity of Statistical Tests on Microbiome Data

Lingjing Jiang, Amnon Amir, Ruth Heller, Ery Arias-Castro, Rob Knight
University of California, San Diego, USA

The analysis of high-dimensional microbiome data often involves performing simultaneous hypothesis tests on each of hundreds or thousands of bacteria in order to detect interesting candidates for further investigation. Classical multiple hypothesis testing methods utilize the false discovery rate (FDR) to control the expected proportion of erroneous rejections among all rejections, with the goal of identifying as many significant findings as possible, while incurring a relatively low proportion of false positives. Due to the discreteness of test statistics in microbiome data, usually caused by the excessive 0s and dominance of a small number of highly abundant bacteria, the commonly used Benjamini Hochberg FDR procedure is often over-conservative, leading to loss of power in detecting significant bacteria. We introduce the Discrete FDR method, which uses permutation-based FDR estimation to utilize the discreteness of the test statistics. Using simulations and real datasets, we demonstrate that it yields increased sensitivity of statistical tests compared to Benjamini Hochberg procedure under the same FDR control level, thus enabling the detection of a larger number of significant findings in sparse and noisy microbiome data.

Email: serene1030@gmail.com

ThAM1-2 T3

Use of a Discrete False Discovery Rate Method for Flagging Potential Safety Signals in Clinical Trials

Joseph Heyse
Merck Research Laboratories, USA

Almost all clinical trials are designed with the objective of evaluating the efficacy of the pharmaceutical, biological, or vaccine product. Evaluating safety is recognized as a primary objective and study teams use rigorous methods to collect, process, and analyze adverse experiences reported by the study participants. Data are carefully cataloged and summarized using standard coding dictionaries such as MedDRA. The underlying problem has been clearly identified as the potential for too many false positive safety findings if the multiplicity problem is ignored. Mehrotra and Heyse (2004) proposed the use of false discovery rate (FDR) control for clinical adverse event data, and an improvement of that method was proposed by Mehrotra and Adewale (2012). This talk will discuss the multiplicity problem as it relates to clinical adverse event data and the suitability of FDR control for this application. In addition, a fully discrete adaption of the Benjamini-Hochberg FDR control method is proposed as a more powerful alternative.

Email: joseph_heyse@merck.com

ThAM1-2 T4

Procedures Controlling the FWER for Discrete Data

Li He, Joseph Heyse
Merck Research Laboratories, USA

In many applications where it is necessary to test multiple hypotheses simultaneously, the data encountered are categorical. In such cases, it is important for multiplicity adjustments to take into account the discreteness and heterogeneity of the null distribution of the test statistics, to assure that the procedure is not overly conservative. When the number of hypotheses is small, it is possible to obtain the complete joint null distribution of the test statistics, which can be used to derive exact multiple testing procedures that simultaneously adjust for the discreteness, heterogeneity and dependency among the test statistics. In this paper, these ideas are explored and we derive exact multiple testing procedures that control the familywise error rate (FWER) using the joint null distribution of the discrete test statistic. Performances of the proposed procedures are investigated through simulation studies and real data applications.

Email: li.he@merck.com

ThAM1-2 T5

Exact Approach for Post Hoc Analysis of a Chi-Squared Test

Guogen Shan, Shawn Gerstenberger
University of Nevada Las Vegas, USA

A chi-squared test is often used for testing independence between two factors with nominal levels. When the null hypothesis of independence between two factors is rejected, we are often left wondering where does the significance come from. Cell residuals, including standardized residuals and adjusted residuals, are traditionally used in testing for cell significance, which is often known as a post hoc test after a statistically significant chi-squared test. In practice, the limiting distributions of these residuals are utilized for statistical inference. However, they may lead to different conclusions based on the calculated p-values, and their p-values could be over- or under-estimated due to the unsatisfactory perform of asymptotic approaches with regards to type I error control. Therefore, we propose new exact p-values based on three commonly used test statistics to order the sample space. We theoretically prove that the proposed new exact p-values based on these test statistics are the same. Based on our extensive simulation studies, we show that the existing asymptotic approach based on adjusted residual is often more likely to reject the null hypothesis as compared to the exact approach. We would recommend the proposed exact p-value for use in practice as a valuable post hoc analysis technique for chi-squared analysis.

Email: guogen.shan@unlv.edu

ThAM1-3 T1

Implementing Monte Carlo Tests with Multiple Thresholds

Georg Hahn, Axel Gandy, Dong Ding
Imperial College London, UK

Software packages usually report the significance of statistical tests using p-values. We are interested in computing the significance of a hypothesis H with respect to several thresholds simultaneously with the caveat that the p-value p corresponding to H is unknown and can only be approximated using Monte Carlo simulation. Instead of considering a set of thresholds, this talk presents a more general construction which allows to compute a decision of p with respect to user-specified intervals (called "p value buckets"): Whereas non-overlapping buckets lead to classical decisions in expected infinite runtime, suitably chosen overlapping buckets allow guaranteed decisions in finite time which are reported in a new fashion that extends the widespread */**/*** significance notation.

Email: g.hahn11@ic.ac.uk

ThAM1-3 T2

Estimating the Proportion of True Null Hypotheses Under Dependency

Thorsten Dickhaus, Andre Neumann, Taras Bodnar
University of Bremen, Germany

It is a well known result in multiple hypothesis testing that the proportion π_0 of true null hypotheses is not identified under general dependencies. However, it is possible to estimate π_0 if structural information about the dependency structure among the test statistics or p-values, respectively, is available. We demonstrate these points, and propose a marginal parametric bootstrap method. A pseudo-sample of bootstrap p-values is generated, which still carry information about π_0, but behave like realizations of stochastically independent random variables. Theoretical properties of resulting estimation procedures for π_0 are analyzed and their usage is illustrated on synthetic and real data

Email: dickhaus@uni-bremen.de

ThAM1-3 T3

Permutation-Based Simultaneous Confidence Bounds for the False Discovery Proportion

Jesse Hemerik, Aldo Solari, Jelle Goeman
Leiden University Medical Center, Netherlands

When many hypotheses are tested, interest is often in ensuring that the proportion of false discoveries (FDP) is small with high confidence. We construct confidence upper bounds for the FDP, which are simultaneous over all rejection cut-offs. In particular this allows the user to select a set of hypotheses such that the FDP lies below some constant with high confidence. Our methods use permutations to account for the dependence structure in the data. So far only Meinshausen provided an exact, permutation-based and computationally feasible method for uniform FDP bounds. We improve this procedure by embedding it within a closed testing framework. Further, we provide a generalization of the method. It lets the user specify a set from which the confidence envelope is selected. This gives the user more freedom in determining the properties of the method. For example, the user can prioritize certain rejection cut-offs, obtaining better FDP bounds for these cut-offs at the cost of larger bounds for other cut-offs. Interestingly, several existing permutation methods, such as SAM and Westfall and Young's maxT method, are obtained as special cases. The different procedures in this paper are compared using both simulated and real data.

Email: j.b.a.hemerik@lumc.nl

ThAM1-3 T4

Significant Pattern Mining on Graphs

Mahito Sugiyama
Osaka University, Japan

I will review techniques of significant pattern mining from graph databases. A representative application is significant subgraph mining, where the objective is to enumerate statistically significantly enriched subgraphs in one of two collections of graphs while correcting for multiple testing. I will show that pruning untestable subgraphs using Tarone's testability trick is the key to solve the problem.

Email: mahito@nii.ac.jp

ThAM1-3 T5

Controlling FWER and FDR in Emerging Pattern Mining

Junpei Komiyama, Masakazu Ishihata, Hiroki Arimura, Takashi Nishibayashi, Shinichi Minato
University of Tokyo, Japan

Emerging patterns are the patterns whose support significantly changes between two databases. The emerging patterns are useful in many real-world tasks, such as medical tasks, time-series analyses, and classification in machine learning. We study the problem of listing emerging patterns with a multiple testing guarantee. Recently, Terada et al. proposed Limitless Arity Multiple-testing Procedure (LAMP), a method that controls the family-wise error rate (FWER) in statistically significant associations. LAMP is able to increase its statistical power by reducing "untestable" hypotheses. Still, FWER is restrictive, and as a result its statistical power is inherently unsatisfying when the number of patterns is large. On the other hand, the false discovery rate (FDR) is less restrictive than FWER, and thus controlling FDR can yield a more larger number of significant patterns. We propose two emerging pattern mining methods: the first one controls FWER, and the second one controls FDR. The effectiveness of the methods are verified by computer simulations with real world datasets.

Email: jkomiyama@tkl.iis.u-tokyo.ac.jp

ThAM2-1 T1

Nonparametric Inference Following Adaptive Designs with Sample Size Reassessment

Florian Klinglmueller, Martin Posch, Livio Finos
AGES-Austrian Agency for Health & Food Safety, Austria

Type I error control following adaptive designs can be achieved by using hypothesis tests based on combination functions or the conditional error rate principle. While previous work has focused on parametric testing procedures, we investigate nonparametric adaptive tests and derive an adaptive randomization test based on the conditional error rate principle. The proposed procedures allow for sample size increases based on the unblinded interim data. To guarantee type I error rate control no specific sample size reassessment procedure has to be pre-specified. We suggest efficient rules for adaptive sample size extension, that on average require fewer samples to achieve the same power as corresponding fixed sample designs. We show that the proposed tests are robust in terms of power for a wide variety of outcome distributions and outperform existing tests for adaptive trials, especially when sample sizes are small.

Email: florian.klinglmueller@meduniwien.ac.at

ThAM2-1 T2

Correcting for Selection Bias in Adaptive Two-Stage Designs

David Robertson, Toby Prevost, Jack Bowden
MRC Biostatistics Unit, University of Cambridge, UK

The problem of selection bias has long been recognized in the analysis of two-stage clinical trials, where promising candidate treatments are selected in stage 1 for confirmatory analysis in stage 2. Specifically, a treatment has to perform 'well' in stage 1 in order to proceed to stage 2, which can lead to overly-optimistic estimates at the end of the trial. To efficiently correct for bias, the uniformly minimum variance conditionally unbiased estimator (UMVCUE) has been proposed for a variety of trial settings, but where the parameter estimates are assumed to be independent. We relax this assumption and derive the UMCVUE in the multivariate normal setting with an arbitrary known correlation structure. A key application is the estimation of treatment effects in adaptive seamless phase II/III clinical trials. Methods for bias adjustment developed thus far have made restrictive assumptions about the design and selection rules followed. Our framework allows for the precision of the treatment arm estimates to take arbitrary values; can be utilized for all treatments that are taken forward to phase III; and is applicable when the decision to select or drop treatment arms is driven by a multiplicity-adjusted hypothesis testing procedure.

Email: david.robertson@mrc-bsu.cam.ac.uk

ThAM2-1 T3

Design and Monitoring of Multi-Arm Multi-Stage Clinical Trials

Pranab Ghosh, Cyrus Mehta
Cytel Inc, Boston University, USA

The statistical methodology for two arm group sequential clinical trials has been available for at least 35 years. The generalization to adaptive two-arm group sequential designs became available only in the last decade thanks to seminal papers by Lehmacher and Wassmer (1999), Cui, Hung and Wang (1999) and Muller and Schafer (2000). The very next stage of development is the generalization of these methods to multi-arm multi-stage (MAMS) group sequential trials. The statistical methodology already developed for the two-arm case can, in principle, be extended to MAMS designs. In practice, however, the formidable computational problems that must be overcome have inhibited making these methods accessible for realistic designs. We will discuss our recent work on overcoming these computational hurdles and will demonstrate the use of these methods for adaptive clinical trials.

Email: pranabg@bu.edu

ThAM2-1 T4

Analytical and Empirical Comparison of MAMS and P-Value Combination Approaches for Adaptive Designs

Cyrus Mehta, Pranab Ghosh, Lingyun Liu
Cytel Inc. USA

Multi-arm multi-stage (MAMS) designs are designs that compare several intervention arms to a common control arm in a randomized clinical trial with one or more interim analyses at which arms can be terminated either for futility or overwhelming efficacy. There are two approaches for constructing such designs. The p-value combination approach, with closed testing to ensure strong control of type-1 error, is the method that is most frequently used. Recently, however, there has been a great deal of interest in the extension of group sequential methods from two arm trials to multi-arm trials with stopping boundaries derived from error spending functions. In this presentation we will discuss the methodological difference between the two approaches and compare their operating characteristics in various settings including adaptive sample size re-estimation.

Email: mehta@cytel.com

ThAM2-2 T1

Optimal Data-Driven Weighting Procedure with Grouped Hypotheses

Guillermo Durand
Universite Pierre et Marie Curie, France

The Benjamini-Hochberg (BH) procedure is a well-known FDR-controlling procedure whose power can be improved by weighting the p-values. This paper provides an optimal way of doing this by defining a new fully computable weighted step-up procedure. The FDR control is proved and the power optimality is achieved in a certain sense, in the context of grouped hypotheses. This study is based on the works of Roquain and van De Wiel (2009) who present an oracle optimal weighted procedure, and Zhao and Zhang (2014) who also provide a data-driven weighted procedure, but without optimality.

Email: guillermo.durand@upmc.fr

ThAM2-2 T2

A General Convex Framework for Multiple Testing with Prior Information

Edgar Dobriban
University of Pennsylvania, USA

Using prior information may improve power in frequentist multiple testing. P-value weighting is a promising methodology where each test is conducted at a different level, using critical values based on independent prior data. However, existing methods are limited, and do not allow the user to specify properties of the weights that are desired in practice, such as boundedness, or monotonicity in the strength of prior evidence. Here we develop a general framework for p-value weighting based on convex optimization. This allows flexible constraints and leads to a variety of new methods, such as bounded and monotone weights, stratified weights, and smooth weights. It also recovers several existing heuristics. Finally we focus on the promising special case of bounded monotone weights. These are appealing as they increase with the strength of prior evidence, and stable because they are within pre-specified bounds. We show they have good empirical power in the analysis of genome-wide association studies.

Email: dobriban@wharton.upenn.edu

ThAM2-2 T3

A Unified Framework for Weighted Parametric Multiple Test Procedures

Dong Xi, Ekkehard Glimm, Willi Maurer, Frank Bretz
Novartis, USA

We describe a general framework for weighted parametric multiple test procedures based on the closure principle. We utilize general weighting strategies that can reflect complex study objectives and include many procedures in the literature as special cases. The proposed weighted parametric tests bridge the gap between rejection rules using either adjusted significance levels or adjusted p-values. This connection is made by allowing intersection hypotheses of the underlying closed test procedure to be tested at level smaller than α. This may be also necessary to take certain study situations into account. For such cases we introduce a subclass of exact α-level parametric tests which satisfy the consonance property. When the correlation is known only for certain subsets of the test statistics, a new procedure is proposed to fully utilize this knowledge within each subset. We illustrate the proposed weighted parametric tests using a clinical trial example.

Email: dong.xi@novartis.com

ThAM2-2 T4

Conditionalized Testing: Improvement of Multiple Testing Methods When Testing Inflated p-Values

Jakub Pecanka, Jules Ellis, Jelle Goeman
Leiden University Medical Center, Netherlands

Many multiple hypothesis testing scenarios lead to inflated representation of large p-values, i.e. those with values near 1. This occurs for instance when interval null hypotheses are tested. A common practice in such situations is to simply remove the very large p-values (those above a fixed threshold λ) and proceed with the analysis as if the large p-values were never observed. However, this is an anti-conservative strategy, which inflates the chance of false findings (as measured by either FWER or FDR). We show that for many multiple testing procedures (e.g. Bonferroni, Holm, Hommel, Benjamini-Hochberg methods) under many scenarios the anti-conservativeness can be cured by employing a strategy called emphconditionalized testing, where in addition to removing the p-values above λ the user also re-scales the remaining p-values by λ and subsequently applies the selected multiple testing procedure to the resulting set of p-values. Crucially, the scenarios where conditionalization leads to valid testing procedures include setups with independent p-values (for any distribution of test statistics) and positively dependent p-values (for normally distributed test statistics) provided that the p-values marginally dominate the uniform distribution. We provide both theoretical results and numerical illustrations.

Email: j.pecanka@lumc.nl

ThAM2-2 T5

Adaptive Multiple Hypothesis Testing for Complex Networks and High Dimensional Data

Djalel-Eddine Meskaldji
Ecole polytechnique federale de Lausanne EPFL, Switzerland

Virtual and real relationships between variables can be represented by networks consisting of nodes and edges. A variety of measures can be used to assess the topological structure of networks at different scales, from local to global. However, testing local hypotheses (e.g., at the node/edge level) involves the multiple testing (MT) problem. In this work we show how to exploit the structure of the network in order to improve the power of rejecting false hypotheses while controlling the rate of rejecting true hypotheses (false positives, FP). We will show that our method controls the sFDR defined by E(FP/s(R)), under positive dependence, where R is the total number of rejections and s is a non-decreasing function. The control of the sFDR covers most existing metrics such as FWER, PFER, FDR and FER, and gives more flexibility in the choice of the type I error metric to control. The new method is based on converting the p-value correction to a weighting estimation problem. We show how to choose optimal weights in order to maximise power, also evaluated with different metrics. We also use the concept of borrowing strength in order to have an accurate estimation of the weights, which reduces the variance of FP. We show by means of spatial and network simulated and real data, the gain that could be achieved when considering dependency with our method.

Email: djalel.meskaldji@epfl.ch

ThAM2-3 T1

Selective Inference on a Tree of Hypotheses: New Error Rates and Controlling Strategies

Marina Bogomolov, Christine Burns Peterson, Yoav Benjamini, Chiara Sabatti
Technion, Israel Institute of Technology, Israel

In many complex multiple-testing problems the hypotheses are divided into families which are organized hierarchically in a tree structure. Each family is selected and tested only if all its ancestor hypotheses are rejected. We address the situation where the p-values for parent hypotheses are dependent on the p-values for the hypotheses within the families they index. We formulate a general class of error rates addressing selective inference on families which are organized hierarchically in a tree structure and propose a powerful hierarchical testing procedure with a guaranteed control of such error rates.

Email: marina.bogomolov1@gmail.com

ThAM2-3 T2

Private False Discovery Rate Control and Robustness of the Benjamini-Hochberg Procedure

Weijie Su, Cynthia Dwork, Li Zhang
University of Pennsylvania, USA

We provide the first differentially private algorithms for controlling the false discovery rate (FDR) in multiple hypothesis testing. Our general approach is to adapt a well-known variant of the Benjamini-Hochberg procedure (BHq), making each step differentially private. This destroys the classical proof of FDR control. To prove FDR control of our method, we develop a new proof of the original (non-private) BHq algorithm and its robust variants - a proof requiring only the assumption that the true null test statistics are independent, allowing for arbitrary correlations between the true nulls and false nulls. This assumption is fairly weak compared to those previously shown in the vast literature on this topic, and explains in part the empirical robustness of BHq.

Email: suweijie444@gmail.com

ThAM2-3 T3

A Unified Treatment of Multiple Testing with Prior Knowledge

Aaditya Ramdas, Rina F. Barber, Michael I. Jordan, Martin J. Wainwright
University of California, Berkeley, USA

A significant literature studies ways of employing prior knowledge to improve power and precision of multiple testing procedures. Some common forms of prior knowledge may include (a) a priori beliefs about which hypotheses are null, modeled by non-uniform prior weights; (b) differing importance of hypotheses, modeled by differing penalties for false discoveries; (c) multiple arbitrary partitions of the hypotheses into known (possibly overlapping) groups, indicating (dis)similarity of hypotheses; and (d) knowledge of independence, positive dependence or arbitrary dependence between hypotheses or groups, allowing for more aggressive or conservative procedures. We unify a number of existing procedures, generalize the conditions under which they are known to work, and simplify their proofs of FDR control. Then, we present an algorithmic framework for global null testing and false discovery rate (FDR) control that allows the scientist to incorporate all four types of prior knowledge (a)_(d) simultaneously, recovering a wide variety of common algorithms as special cases.

Email: aramdas@berkeley.edu

ThAM2-3 T4

Optimal Rates and Tradeoffs in Multiple Testing

Maxim Rabinovich, Aaditya Ramdas, Martin Wainwright, Michael Jordan
University of California, Berkeley, USA

Multiple hypothesis testing has become a central problem in both applied and theoretical statistics, with a number of methodologies proposed to achieve good performance under a variety of metrics, including familywise error rate (FWER) and false discovery rate (FDR). Despite particular interest in FDR and the corresponding measure of power known as the false negative rate (FNR), the field has not achieved a detailed understanding of fundamental lower bounds and tradeoffs between FDR and FNR. In this paper, we establish an on-asymptotic tradeoff between FNR and FDR in a generalized Gaussian sequence model previously studied in the multiple testing literature. We base our analysis on a novel yet simple comparison principle that allows us to sidestep the difficult asymptotics of empirical CDFs. Our methodology is flexible enough to permit analysis of regimes where the problem parameters vary with n, including relatively dense regime with n/logn or αn signals. Perhaps surprisingly, we find that across all regimes, the popular BH and BC algorithms are optimal up to constants, though their rate may become subpolynomial in n.

Email: rabinovich@eecs.berkeley.edu

ThPM1-1 T1

Flexible Statistical Approaches for Biosimilar Development

Pantelis Vlachos
Cytel, Inc., USA

Considerable interest has grown among pharmaceutical and other medical product developers in adaptive clinical trials, in which data collected during the course of a trial affects ongoing conduct or analysis of the trial. Following the release of the FDA draft Guidance document on adaptive design clinical trials in early 2010, expectations of an increase in regulatory submissions involving adaptive design features, particularly for confirmatory trials, were high. In this talk we review flexible such approaches for biosimilar development. We will summarize the major statistical methods by clinical development phase. Challenges of adaptive trial implementations will be discussed and recommendations will be provided.

Email: pantelis.vlachos@cytel.com

ThPM1-1 T2

Unblinded Sample Size Re-Estimation in Bioequivalence Trials with Small Sample Sizes

Sam Hsiao, Lingyun Liu, Romeo Maciuca
Cytel, Inc., USA

We consider a framework for inference in adaptive bioequivalence trials with unblinded sample size re-estimation (SSR). If the sample size is small and the variance unknown, as is often the case in bioequivalence trials, using boundaries derived under the assumption of a normally distributed test statistic may lead to type I error inflation. While this problem can be overcome with p-value combination methods, these approaches generally do not directly provide confidence intervals for the geometric mean ratio on the scale of the original pharmacokinetic endpoint. We consider an approach that involves pre-specifying a range of final sample sizes to allow some flexibility in the SSR procedure, yet uses pre-defined constant boundaries based on a "piecewise t-distribution" to derive repeated confidence intervals (RCIs) for the treatment effect. The RCIs have guaranteed coverage, and can be used for inference and clinical interpretation in the same way that conventional two-sided confidence intervals are typically used when applying the two one-sided testing (TOST) procedure.

Email: sam.hsiao@cytel.com

ThPM1-1 T3

Understanding Biosimilarity by Totality of the Evidence: What should Statisticians Know

Yushi Liu, Jason Hsu
Eli Lilly and Company, USA

Nowadays, more and more generic drugs are available to patients as an alternative. As chemical compound structures are simpler to characterize, biologics are much more challenging. In FDA's regulatory guidance, the totality-of-the-evidence approach is recommended to demonstrate biosimilarity between the proposed product and a reference product. For this approach, the function and structural characterization as well as clinical safety and efficacy are important but this is sometimes out of the knowledge scope for statisticians. To facilitate the role of statisticians in testing biosimilarity, we will give an introduction on testing interchangeability and bioequivalence and analytical techniques in the biosimilar scenarios.

Email: liu_yushi@lilly.com

ThPM1-1 T4

A Distribution-Free Consistency Adjusted Stepwise Testing Procedure

Jaclyn McTague, Dror Rom
Prosoft Clinical, USA

Strategies for designing confirmatory clinical studies used to obtain efficacy claims from regulatory agencies vary, but often include a fallback option in case the primary endpoint does not meet its projected efficacy. Alosh and Huque (2010) have discussed a problem that is sometimes encountered when two endpoints that can fully characterize the treatment effect on their own, and are used in a testing scheme to support approval of a treatment, produce inconsistent results, leading to a problem of interpretation of the findings. They make an appealing argument that important endpoints should produce a minimum degree of agreement. They develop a range of procedures that utilize an internal measure of consistency between the results of the two endpoints used in the testing scheme. The exact calculation of the critical points in their procedures require that the joint distribution of the endpoints to be bivariate normal with a known correlation. In this paper, we develop a distribution-free analogue of the Alosh and Huque (2010) procedure which we show to have type-1 error control under mild regulatory assumptions, exhibited for example in the bivariate normal and bivariate t having either a positive or a negative correlation.

Email: J.McTague@Prosoftclinical.com

ThPM1-2 T1

Confidence Regions for Treatment Effects in Biomarker Stratified Designs

Thomas Jaki, Fang Wan, Cornelia Kunz
Lancaster University, UK

Subgroup analysis has important applications in the analysis of controlled clinical trials. Sometimes the result of the overall group fails to demonstrate that the new treatment is better than the control therapy, but for a subgroup of patients the treatment benefit may exist; or sometimes the new treatment is better for the overall group but not for a subgroup. Hence we are interested in constructing simultaneous confidence regions for the difference of the treatment effects in subgroup(s) and the whole population. Subgroups are usually formed on the basis of a predictive biomarker such as age, sex or some genetic marker. While, for example age can be detected precisely, it is often only possible to detect the biomarker status with a certain probability. Because patients detected with a positive or negative biomarker may not be truly biomarker positive or negative, responses in the subgroups depend on the treatment therapy as well as on the sensitivity and specificity of the assay used in detecting the biomarkers. In this talk we show how (approximate) simultaneous confidence intervals and confidence ellipsoid for the treatment effects in subgroups can be found for biomarker stratified clinical trials using a normal framework. We show that these intervals maintain the nominal confidence level via simulations.

Email: jaki.thomas@gmail.com

ThPM1-2 T2

Confident Inference for SNP Effects on Treatment Efficacy

Ying Ding
University of Pittsburgh, USA

Our research is for finding SNPs that are predictive of treatment efficacy, to decide which subgroup (with enhanced treatment efficacy) to target in drug development. Testing SNPs for lack of association with treatment outcome is inherently challenging, because any linkage disequilibrium between a non-causal SNP with a causal SNP, however small, makes the zero-null (no-association) hypothesis technically false. Control of Type I error rate in testing such null hypotheses are therefore difficult to interpret. We propose a completely different formulation. For each SNP, we provide simultaneous confidence intervals directed toward detecting possible dominant, recessive, or additive effects. Across the SNPs, we control the expected number of SNPs with at least one false confidence interval coverage while taking the correlation among SNPs into account. Since our confidence intervals are constructed based on pivotal statistics, the false coverage control is guaranteed to be exact and unaffected by parameter values (whether zero or non-zero). Our method is applicable to the therapeutic areas of Diabetes and Alzheimer's diseases, as a step toward condent targeting of patient subgroups in a tailored drug development process.

Email: yingding@pitt.edu

ThPM1-2 T3

Statistical Issues in Subgroup Discovery Using Permutation Testing

Siyoen Kil
LSK Global PS, Korea

The goal of the subgroup analyses in clinical trials is to quantify the heterogeneity of treatment effect across subpopulation. Identifying the targeted subgroup where the treatment effect is enhanced compare to the complement subgroup can benefit both patients and drug developers. Subgroup analyses for drug development should involve multiple testing. In the situation in which multiple hypotheses should be tested, permutation testing is very prevalent. For example, Jiang et al.(2007) discuss permuting treatment labels for the establishment and validation of a cut point for pre-defined subgroup. Freidlin et al.(2010) and Lipkovich et al.(2011) use permutation to explore candidate biomarkers that may define the target subgroup. The draft guidance of FDA, "Adaptive Design Clinical Trials for Drugs and Biologics" refers Freidlin and Simon (2005) that uses permutation to test a gene expression signature for sensitive patients. Permutation testing is very tempting because practitioners might expect the correlation structure among statistics for each null hypothesis to be retained by the permutation while still controlling the Type-I error with better power. Retaining the correlation structure can give the great advantage of nominal (not conservative) testing result, especially for highly correlated statistics. Permutation, however, does not generally produce correct reference distributions. Huang et al. (2006) and Xu and Hsu (2007) note that permutation is only appropriate for multiple testing when the marginal null hypotheses determine the joint distribution (MDJ condition) of the test statistics. Calian et al. (2008) and Kaizar et al. (2011) demonstrate that multiple testing for prognostic biomarkers without MDJ condition does not always control the family-wise error rate(FWER). In the presentation, it will be discussed that the permutation testing for biomarker-treatment interaction for binary response to find a target subgroup does not always control the Type-I error even with only one subgroup classifier. Underlying idea for testing this predictive biomarker is the same as MDJ in that usually permuting one label (biomarker labels or treatment labels) makes reference distribution with zero marginal effect (this means testing neither prognostic nor predictive biomarker which is not the one we want to test) which is not always true. I will also discuss a valid permutation method that corresponds to the hypothesis of interest (specifically, homogeneous treatment effect through subgroup).

Email: siyoenk@gmail.com

ThPM1-2 T4

A Case Study in Precision Medicine: Rilpivirine Versus Efavirenz for Treatment-Naive HIV Patients

Zhiwei Zhang, Wei Liu, Lei Nie, Guoxing Soon
University of California, Riverside

Rilpivirine and efavirenz are two drugs for treatment-naive adult patients infected with human immunodeficiency virus (HIV). Two randomized clinical trials comparing the two drugs suggested that their relative efficacy may depend on baseline viral load and CD4 cell count. Here we estimate individualized treatment regimes that attempt to maximize the virologic response rate or the median of a composite outcome that combines virologic response with change in CD4 cell count (dCD4). To estimate the target quantities for a given treatment regime, we use G-computation, inverse probability weighting (IPW) and augmented IPW methods to deal with censoring and missing data under a monotone coarsening framework. The resulting estimates form the basis for optimization in a class of candidate regimes indexed by a smaller number of parameters. A cross-validation procedure is used to deal with the re-substitution bias in evaluating an optimized treatment regime.

Email: zhiwei.zhang@ucr.edu

ThPM1-3 T1

Sequential Multiple Testing for Biomarker Discovery

Xinping Cui, Hailu Chen
University of California, Riverside, USA

Covariance test (Lockhart et al. 2014) provided a testing procedure to enter variables into a linear model sequentially along a lasso solution path. Using covariance test to select a model with inferential guarantees is equivalent to multiple hypothesis testing setting where the hypotheses are ordered. In this talk, we proposed a sequential multiple hypothesis testing framework, which considers multiple testing within each step and across all steps along the lasso solution. Biomarker discovery in framingham heart study examples are presented as application of the proposed method.

Email: xinping.cui@ucr.edu

ThPM1-3 T2

Unbiased Estimation of Biomarker Panel Performance When Combining Training and Testing Data in a Group Sequential Design

Nabihah Tayob, Kim-Anh Do, Ziding Feng
MD Anderson Cancer Center, USA

Motivated by an ongoing study to develop a screening test able to identify patients with undiagnosed Sjogren's Syndrome in a symptomatic population, we propose methodology to combine multiple biomarkers and evaluate their performance in a two-stage group sequential design that proceeds as follows: biomarker data is collected from first stage samples; the biomarker panel is built and evaluated; if the panel meets pre-specified performance criteria the study continues to the second stage and the remaining samples are assayed. The design allows us to conserve valuable specimens in the case of inadequate biomarker panel performance. We propose a nonparametric conditional resampling algorithm that uses all the study data to provide unbiased estimates of the biomarker combination rule and the sensitivity of the panel corresponding to specificity of 1-t on the receiver operating characteristic curve (ROC). The Copas and Corbett (2002) correction, for bias resulting from using the same data to derive the combination rule and estimate the ROC, was also evaluated and an improved version was incorporated. An extensive simulation study was conducted to evaluate finite sample performance and propose guidelines for designing studies of this type. The methods were implemented in the National Cancer Institutes Early Detection Network Urinary PCA3 Evaluation Trial.

Email: ntayob@mdanderson.org

ThPM1-3 T3

Elucidating Issues of Multiplicity that Arise with Clinical Trial Designs for Precision Medicine

Brian Hobbs, Nan Chen
MD Anderson Cancer Center, USA

Human diseases can be intrinsically heterogeneous with respect to their pathogenesis among patient populations and in some contexts vary in composition among multiple locations within a patient. Owing to pharmacogenetic diversities, a particular therapeutic strategy may yield very different outcomes among patients with similar diagnoses. In the oncology setting, wherein the majority of definitive comparisons in phase III fail to demonstrate the hypothesized extent of benefit for new treatment strategies for solid tumors, it is widely accepted that the relative utility of a given therapy is determined by the confluence of a patient's particular clinical prognosis as well as the tumor's particular molecular composition. In fact, drug development strategies devised to characterize cohort-averaged treatment benefits have largely failed in oncology with only 34% of confirmatory phase III trials yielded a significant result from 2003 to 2010 and final market approval achieved for only 13% of the cancer drugs that initiated phase I between 1993 and 2004. Implicit to the concept of precision medicine is heterogeneity of treatment benefit among patients and patient sub-populations. Recent advances in design methodology used in oncology endeavor to study many agent-and-target combinations in parallel. Multiplicities arise with the analyses of these trials for which clear guidelines from the statistical community have yet to be established. My presentation is intended to elucidate multiplicity issues for designs used in precision medicine contexts and discuss potential avenues for establishing guidelines.

Email: bphobbs@mdanderson.org

ThPM1-3 T4

An Application of FDR to Billions of Hypothesis Testing to Identify Expression Quantitative Trait Loci in Genome Wide Association Studies

Irina Dinu, Fahimeh Moradi, Elham Khodayari-Moez
University of Alberta, Canada

Introduction: Genome wide association studies (GWAS) have been widely used in recent years to identify new information on genetic variants which are associated with complex trait in many diseases. Advances in identifying the Single nucleotide polymorphisms (SNPs) facilitate the study of etiologies of common disorders including cancers, inflammatory bowel diseases (IBD) and colorectal cancer. However, the known SNPs are not sufficient to explain the heritability associated with traits. Variations in gene expression demonstrate that transcript levels of many RNAs behave as heritable quantitative traits. Studying the genetics of gene expression can provide additional power to the roles of GWAS variants. Expression quantitative trait loci (eQTL) mapping links the genome-wide SNPs with RNA expression.
Objective: Our objective is to identify an efficient, statistically sound and user friendly method for analysis of eQTL studies.
Methods: In this study, we performed expression quantitative trait loci (eQTL) analysis using the Matrix eQTL R package. This technique implements matrix covariance calculation and efficiently runs linear regression analysis. The statistical test determines the association between SNP and gene expression, where the null hypothesis is no association between genotype and phenotypes. In eQTL mapping, the regulative variants are classified as cis and trans, the definition depending on the physical distance between a gene and transcript. A certain genomic distance (e.g. 1 Mb) is defined as the maximum distance at which cis or trans regulatory elements can be located from the gene they regulate. False discovery rate (FDR) is used to identify significant cis and trans eQTL for multiple testing corrections.
Results: We applied matrix eQTL to a real data set consisting of 730,256 SNP and 33,298 RNA for 173 samples. SNPs with minor allele frequency (MAF) less than 0.05 and those violating the Hardy-Weinberg equilibrium (HWE), were excluded from the study. In this study, 15,408 cis eQTL and 27,562 trans eQTL are identified at a FDR less than 0.05, corresponding to p value thresholds of 8e-5 and 1e-8, respectively.
Conclusion: We found out that matrix eQTL is a computationally efficient and user friendly method for analysis of eQTL studies. The results provide insight into the genomic architecture of gene regulation in inflammatory bowel disease (IBD).

Email: idinu@ualberta.ca

FAM1-1 T1

Adaptive Sequential Model Selection

William Fithian, Jonathan Taylor, Robert Tibshirani, Ryan Tibshirani
University of California, Berkeley, USA

Many model selection algorithms produce a path of fits specifying a sequence of increasingly complex models. Given such a sequence and the data used to produce them, we consider the problem of choosing the least complex model that is not falsified by the data. Extending the selected-model tests of Fithian et al. (2014), we construct p-values for each step in the path which account for the adaptive selection of the model path using the data. In the case of linear regression, we propose two specific tests, the max-t test for forward stepwise regression (generalizing a proposal of Buja and Brown (2014)), and the next-entry test for the lasso. These tests improve on the power of the saturated-model test of Tibshirani et al. (2014), sometimes dramatically. In addition, our framework extends beyond linear regression to a much more general class of parametric and nonparametric model selection problems. To select a model, we can feed our single-step p-values as inputs into sequential stopping rules such as those proposed by G'Sell et al. (2013) and Li and Barber (2015), achieving control of the familywise error rate or false discovery rate (FDR) as desired. The FDR-controlling rules require the null p-values to be independent of each other and of the non-null p-values, a condition not satisfied by the saturated-model p-values of Tibshirani et al. (2014). We derive intuitive and general sufficient conditions for independence, and show that our proposed constructions yield independent p-values.

Email: wfithian@berkeley.edu

FAM1-1 T2

Bootstrap Inference After Using Multiple Queries for Model Selection

Jelena Markovic, Jonathan Taylor
Stanford University, USA

Recently, Tian Harris and Taylor (2015) developed a selective inference approach with a randomized response, demonstrating that such randomization provides a way to tradeoff the information used for model selection and the leftover information, used for inference about model parameters. They constructed an asymptotically pivotal test statistic using a CLT that holds without selection. Adjusting the resulting Gaussian limit for selection via a selective CLT allows for selective inference in non-parametric settings. In this work, we provide a refinement of their selective CLT result, most notably we relax their local alternatives assumption. Under some regularity assumptions on the density of the randomization, including heavier tails than Gaussian satisfied by e.g. logistic distribution, we prove the selective CLT holds without any assumptions on the underlying parameter, allowing for rare selection events. We also show that under the local alternatives assumption on the parameter, selective CLT holds for Gaussian randomization as well, though the quantitative results are qualitatively different for the Gaussian randomization as compared to the heavier tailed results. Furthermore, we propose a bootstrap version of this test statistic (bootstrapped pivot). We prove that the bootstrap test statistic is also asymptotically pivotal uniformly across a family of non-parametric distributions. This result can be interpreted as resolving the impossibility results of Leeb and Potscher (2006). We describe a way of using the wild bootstrap and projected Langevin Monte Carlo sampling method to compute a bootstrapped test statistic

and the corresponding confidence intervals valid after selection. As most data analysts will want to try various model selection algorithms when choosing a model, we present a way to construct confidence intervals after multiple views/queries of the data. In this setting, an analyst runs several model selection procedures on the same data and choses a parameter of interest upon seeing the outcomes of all of the model selection procedures. We note that this target of interest need not agree exactly with the results of any model selection procedure - the data analyst can use their own expertise to choose a final parameter of interest but is allowed access to the results of the model selection procedure before choosing their parameter of interest. We construct the selective confidence intervals for the selected parameter using both pivot constructions, plugin CLT and the bootstrap. Finally, we compare our methods to data splitting, in which some portion of the data is used for model selection (stage 1) and the remaining data used for inference (stage 2). In Fithian et al. (2014), it was noted that one can often improve on data splitting by using information leftover after stage 1, though the examples and results were in a parametric setting. In this work, we construct both inferential and sampling tools to reuse the information in the data from the first stage and provide tests with greater power than traditional data splitting. All the computations can be done with any of the examples from Tian Harris et al. (2016), including GLMs, forward-stepwise and marginal screening, and their combination into multiple views framework. We present the implementation results on some of these.

Email: jelenam@stanford.edu

FAM1-1T3

Bayesian Post-Selection Inference in the Linear Model

Snigdha Panigrahi, Asaf Weinstein
Stanford University, USA

In this work, we provide Bayesian inference for a linear model selected after observing the data. Our methodology allows an analyst to choose a generative mechanism post selection and yet be able to provide inference free from any bias from the findings of a selective analysis. Our proposed model consists of a prior and a truncated likelihood, similar to Yekutieli's ideas of adjusting Bayesian inference for selected parameters. The resulting posterior distribution, unlike in the setup usually considered when performing Bayesian variable selection, is affected by the very fact that selection was applied. A major computational challenge in such an approach is the intractability of the truncated likelihood. At the core of our methods is a convex approximation to the truncated likelihood, which facilitates sampling from the (approximate) adjusted posterior distribution. We demonstrate in simulations that employing the proposed approximation results in Bayesian procedures are qualitatively similar to those using the exact truncated likelihood. Replacing the truncated likelihood by its approximation, we can approximate the maximum-likelihood estimate as the MAP estimate corresponding to a constant prior. Our approximation offers a surrogate to full truncated likelihood and hence, is equipped to address frequentist questions that have not been resolved in existing work on exact post-selection inference.

Email: snigdha@stanford.edu

FAM1-1 T4

Selective Sign-Determining Multiple Confidence Intervals with FCR Control

Asaf Weinstein, Daniel Yekutieli
Stanford University, USA

In many areas of science, one observes m independent variables Y_i, each corresponding to a parameter Θ_i in Θ , and the objective is twofold: The primary goal is to detect parameters that belong to each of K disjoint subsets Θ_j subset Θ; The secondary goal is, for each classified parameter, to construct a confidence set with the requirement that the confidence set is a subset of Θ_j if a parameter was declared to belong to Θ_j . This includes the case of constructing compatible confidence sets for parameters of rejected hypotheses after multiple hypothesis testing. We address the problem by proposing a single-stage procedure that constructs selective marginal confidence sets with the required property, while controlling the expected proportion of noncovering confidence sets constructed. Our method is contrasted with some limitations of a conditional approach, taken in our previous work, namely constructing confidence sets with nominal coverage conditional on selection. As a special case we consider the problem of (weak) sign classification of scalar parameters, and propose a configuration of the general procedure that nicely balances a tradeoff between power as a directional decision rule, and length of the constructed sign-determining confidence intervals. Our procedure builds on a new marginal confidence interval designed specifically for the task, and extends the directional step-up procedure of Benjamini and Hochberg.

Email: asafw@stanford.edu

FAM1-2 T1

Group Sequential Designs in Clinical trials with semi-competing risks outcomes

Toshimitsu Hamasaki, Koko Asakura, Scott R Evans, Tomoyuki Sugimoto
National Cerebral and Cardiovascular Center, Japan

Many clinical trials implement group-sequential designs. In some disease areas e.g., oncology or cardiovascular disease, these trials utilize event-time outcomes and are event-driven meaning that interim analyses are performed when a certain number of events have been observed. In such trials, one challenge is how to monitor multiple event-time outcomes in a group-sequential setting as the information fraction for the outcomes may differ at any point in time. We discuss logrank test-based methods for monitoring two event-time outcomes in group-sequential trials that compare two interventions using two time-to-event outcomes. We evaluate two situations: (i) both events are non-composite but one event is fatal, and (ii) one event is composite but the other is fatal and non-composite. We consider several strategies for testing if a test intervention is superior to a control intervention on at least one of the event-time outcomes.

Email: toshi.hamasaki@ncvc.go.jp

FAM1-2 T2

Analysing Multiple Outcomes in Randomised Controlled Trials Using the Multilevel Multivariate Model

Victoria Vickersta_, Gareth Ambler, Rumana Z Omar
University College London, UK

In clinical trials, it is common to have multiple primary outcome measures. These outcome measures are often correlated. Many procedures for analysing multiple outcomes have been suggested. Analysing the outcomes separately, in a univariate framework, is one of the most of the commonly used methods [1]. However, this approach does not make use of the multivariate structure in the data and as such ignores any correlations between the outcomes. To analyse multiple correlated outcomes the multilevel multivariate model (MMM) may be used. The MMM analyses multiple outcomes as repeated measures clustered within individuals [2]. It makes use of correlations among the outcomes which may help when we have missing data. The MMM can handle continuous outcomes, binary outcomes or a mixture of both. If we have responses of different types, for example binary and continuous responses, the observed binary responses are assumed to have an underlying latent normal distribution. Both the continuous and binary responses are mapped onto an underlying multivariate normal distribution. The model can be extended to include three or more outcomes and additional levels of clustering (for example institutions or repeated measures). When evaluating an intervention effect using the MMM either the effect on each outcome can be reported or an overall intervention effect can be reported. The performance of the MMM was investigated using simulation. Specifically, the power to detect true intervention effects was assessed. The MMM results were compared to the power obtained when analysing each outcome separately. A Holm adjustment was used to amend the p-values to control the familywise error rate [2]. Simulation scenarios include varying the number of outcomes, correlation between outcomes and the degree and pattern of missingness. The simulation study shows that the MMM performs better in terms of power when the outcomes are correlated and there are missing data. However, the gains were small except when there was high correlation or high levels of missing data.

[1] Vickersta_ V, Ambler G, King M, Nazareth I, Omar RZ. Are multiple primary outcomes
analysed appropriately in randomised controlled trials? A review. Contemporary clinical trials.
2015 Nov 30;45:8-12.
[2] Goldstein H. Multilevel statistical models. John Wiley & Sons; 2011 Jul 8.

Email: v.vickerstaff@ucl.ac.uk

FAM1-2 T3

Comparison of Novel Approaches in Dose Response Studies

Saswati Saha
University of Bremen, Germany

Characterizing an appropriate dose-response relationship and identifying the right dose in a clinical trial are two main goals of early drug-development . The MCP-Mod is one of the pioneer approaches developed within the last 10 years which combines the modeling techniques with multiple comparison procedures to address the above goals in clinical drug development. The MCP-Mod approach begins with a set of potential dose-response models, tests for a significant dose-response effect using multiple linear contrasts and fits the best model based on modeling techniques. However, there is quite a possibility of model mis-specification in this approach. The non-linear parameters for the candidate models need to be chosen a priori for the multiple contrast tests. This may lead to a loss in power and unreliable model selection as well as model fitting. Motivated by the above shortcomings, we compare MCP-Mod with other dose-finding approaches that a more robust means in dose-response shape detection. In our presentation, we will discuss three state-of-the-art approaches which assume a candidate set of parametric dose-response models and test the null hypothesis of no dose-response trend against the composite alternative that one of the candidate dose-response shapes is true. These approaches do not make prior assumptions about the model parameters and are therefore more robust compared to MCP-Mod approach. Our focus is to compare these approaches with regard to their ability to detect the dose-response trend, potential to select the correct model and accuracy in estimating the minimum effective dose in dose-finding studies in an extensive simulation study.

Email: saswatisaha18@gmail.com

FAM1-2 T4

Comparisons of Efficiency and Robustness of Multiple Testing Procedures in Phase 3 Clinical Trials

Michael Lee, Anjun Cao
Janssen R&D, USA

Multiplicity is common in clinical trials. In the newly released FDA guidance to industry on multiple endpoints in clinical trials, the agency provides an overview of the multiplicity issue in clinical trials and commonly used multiple testing procedures. In practice many factors need to be taken into consideration when pre-specifying primary multiple testing procedures for clinical trials in drug development. Because objectives and strategies vary, there is no one-size-to fit solution. When there is a good knowledge of underlying treatment effect sizes for corresponding endpoints, it is often clear how to identify an efficient multiple testing procedure. Sometimes a simple testing procedure such as a fixed sequence test will work well. When the effect sizes are uncertain, selection of a multiple testing procedure should be based on efficient as well as robustness consideration. Dmitrienko and others proposed evaluation criteria to enable comparisons among multiple testing procedures. In this presentation we will show comparisons of commonly used multiple testing procedures in drug development in terms of efficiency and robustness. Numerical examples will be discussed.

Email: mlee60@its.jnj.com

FAM1-2 T5

Testing Strategy in Phase 3 Trials with Multiple Doses

David Li
Pfizer, USA

There has been a trend to have more than one dose in Phase 3 trials recently, with one dose being targeted dose, and one or two more as backup doses. Multiplicity raised by comparing multiple doses can be addressed by many available approaches. But none of them is completely satisfactory. For example, the concern with Hochberg's approach is that the second step has to use the half of alpha if no rejection occurs at the first step. The concern with the approaches requires consistency (eg, Li, 2001, SiM) is the loss of opportunity to reject some if the consistency criterion fails. This presentation will introduce a new approach which addresses these concerns and is shown to be more powerful than currently available approaches. The idea of the new approach is to test the global null hypothesis using the statistics based on pooled dose and differences of doses, plus inconsequential consistency criteria. When the doses are similar, the statistic based on the pooled dose will likely help reject the null, and when the backup doses are different from the target dose, the statistics based on differences of doses will help reject the null.

Email: david.li1@pfizer.com

FAM1-3 T1

Adaptive Filtering Multiple Testing Procedures for Partial Conjunction Hypotheses

Jingshu Wang, Art B. Owen, Chiara Sabatti
University of Pennsylvania, USA

The partial conjunction (PC) alternative hypothesis $H_1\ ^{r/n}$ stipulates that at least r of n related basic hypotheses are non-null, making it a useful measure of replicability. Motivated by genomic problems we consider a setting with a large number M of partial conjunction null hypotheses to test, based on an n×M matrix of p-values. When r>1 the hypothesis $H_0\ ^{r/n}$ is composite. Validity versus the case with r-1 alternative hypotheses holding can lead to very conservative tests. We develop a filtering approach for $H_0\ ^{r/n}$ based on the M p-values for $H_0\ ^{(r-1)/n}$. This filtering approach has greater power than straightforward PC testing in the multiple testing setting. We prove that it can be used to control the familywise error rate, the per family error rate, and the false discovery rate among M PC tests. In simulations we find that our filtering approach properly controls the FDR while achieving good power. We illustrate application of the method in both microarray data analysis and genome-wide association studies (GWAS).

Email: jingshuw@upenn.edu

FAM1-3 T2

A Novel FWER Controlling Procedure for Data with Reduced Rank Correlation Structure

Xing Qiu
University of Rochester, USA

Recent emergence of high-throughput data such as microarray and RNA-seq data heralds a new era of research in the field of multiple testing. Traditional procedures that were designed to control familywise error rate (FWER) were largely replaced by false discovery rate (FDR) controlling procedures in practice, due to the lack of statistical power of classical FWER controlling procedures. In a recent study, my collaborators and I discovered that if we replace the unrealistic assumption that all hypotheses being tested are independent or weakly dependent by a class of reduced rank correlation structures, we can achieve adequate statistical power and control FWER at a reasonable level simultaneously. The talk is organized in this way: 1. Illustrate the reduced rank correlation structure in real data (RNA-seq and timecourse microbiome data) by SVD/PCA. 2. Establish the theoretical connections between FWER controlling procedures and spherical statistics. 3. Present a novel FWER controlling procedure, rrMTP, that is optimized for such data. 4. Show the superiority of rrMTP as compared with several other procedures in both simulation studies and two real data applications. 5. Discuss future research opportunities, such as designing an FDR controlling procedure optimized for reduced rank structure.

Email: xing.qiu@gmail.com

FAM1-3 T3

An Empirical Bayes Test for Allelic-Imbalance Detection in ChIP-seq

Qi Zhang, Sunduz Keles
University of Nebraska Lincoln, USA

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has enabled discovery of genomic regions enriched with biological signals such as transcription factor binding sites. Allelic-Imbalance detection is a complementary analysis of ChIP-seq data for associating biological signals with single nucleotide polymorphisms (SNPs) and has been successfully used in elucidating functional roles of SNPs. Commonly used statistical approaches for Allelic-Imbalance detection are often based on binomial testing and mixture models, both of which rely on unrealistic assumptions on the distribution of the unobserved allelic probability, and have significant practical shortcomings. We proposed Non-Parametric Binomial (NPBin) test for allelic-imbalance detection and for modeling Binomial data in general. NPBin models the density of the unobserved allelic probability explicitly and non-parametrically, and estimates the empirical null via curve fitting. We demonstrate the advantage of NPBin in terms of the interpretability of the estimated density, and the accuracy in allelic-imbalance detection using simulations and analysis of many ChIP-seq data. We also illustrate the generality of our modeling framework by an effect size estimation problem in the context of Baseball.

Email: qi.zhang@unl.edu

FAM1-3 T4

FDR control on directed acyclic graphs

Jianbo Chen, Aaditya Ramdas, Michael Jordan, Martin Wainwright
University of California, Berkeley

Consider a directed acyclic graph (DAG), where the nodes represent hypotheses that may be tested, and the directed edges specify a partial order in which the hypotheses must be tested. A hypothesis can only be tested and rejected if all its parent hypotheses have already been rejected. While methods exist for family-wise error (FWER) control on DAGs, current False Discovery Rate (FDR) control procedures can only handle trees (where each node has a unique parent). In this paper, we introduce the first algorithm that can provably control FDR on any pre-specified DAG of hypotheses under a variety of dependence between the hypotheses (including independence, positive dependence, arbitrary dependence, and other variants). Our algorithm reduces to known algorithms when the DAG is a line graph or a tree, and reduces to the Benjamini-Hochberg procedure when the DAG has no edges, and is thus a strict generalization of past work. This algorithm can broadly serve as a model for the scientific investigative process, and to demonstrate its practical performance, we conduct a series of simulations and present a concrete application to a gene ontology DAG.

Email: jianbochen@berkeley.edu

FAM1-3 T5

Statistical Analysis for Estimating Multiple Stopped States in Walking Motions

Toshinari Kamakura, Kosuke Okusa
Chuo University, Japan

A Study on walking is called gait analysis and this is based on moving images using video cameras (Okusa and Kamakura, 2011), but in recent years in Japan it has become more difficult to take pictures in public places from the standpoint of privacy protection. We have shifted to research on behavior measurement using radio wave sensor. In research with Doppler sensor, we proposed a statistical model for respiration measurement without contact (Inui,S., Okusa,K., Maeno,K. and Kamakura,T, 2013; Yamamoto,K., Maeno,K. and Kamakura,T. , 2013). In this research, we consider the problems of statistically estimating the speed of pedestrians by using laser sensors and estimating their speed. On the sidewalk, rapid changes in speed sometimes cause accidents, and then a problem of change point of speed, such as stopping and making suspicious behaviors, must be also considered. In this paper, we discuss the change-point problem of speeds, especially focusing on statistically recognizing the stopped state of walking.

Email: kamakura@indsys.chuo-u.ac.jp

FAM2-1 T1

From Higher Criticism and Local Levels of GOF Tests to Confidence Bounds for the Proportion of True Nulls

Helmut Finner, Veronika Gontscharuk, Klaus Strassburger
Institute for Biometrics and Epidemiology, German Diabetes Center (DDZ), Leibniz, Germany

Local levels can be viewed as an interesting characteristic of union-intersection based overall tests. In some recent work, cf. e.g. Gontscharuk, Landwehr, Finner (2015, Biom. J. 57, 159-180; Bernoulli 22, 1331-1363) and Gontscharuk, Finner (2017, Comm. Stat. - Theory Meth. 46, 2332-2342), we studied local levels of union-intersection based goodness of fit (GOF) tests including higher criticism tests, Kolmogorov-Smirnov type tests and Berk-Jones type tests. Local levels indicate regions of high and low sensitivity of such tests. An interesting issue is the asymptotic behavior of local levels for extreme, intermediate and central order statistics. Typically, local levels tend to zero or converge to some positive limit. Thereby, it is impossible that all local levels have a positive limit. Furthermore, by means of suitable local level shape functions we can design new GOF tests with pre-determined local level behavior. We illustrate the local level behavior of various GOF tests by animated plots. In some cases, the finite local level behavior is far away from the asymptotics even for huge sample sizes. Finally, we show how the concept of local levels can be adopted in order to design improved confidence bounds for the proportion of true and false null hypotheses in multiple testing problems with independent p-values.

Email: finner@ddz.uni-duesseldorf.de

FAM2-1 T2

Conditional Error Rate of Decision Made on the Secondary Endpoint

Haiyan Xu, Jason Hsu
Johnson & Johnson, USA

In a decision-making environment with uncertainty, a statistical error rate has meaning only if controlling it controls the rate of incorrect decision. Using the Minimum Effective Dose (MED) setting as an example, Hsu and Berger (1999) showed that controlling Type I error rate under the complete null (the so-called Experiment-wise Error Rate) does not control the rate of incorrect decision, defined to be inferring a MED which is lower than the true MED. The MED setting conveniently illustrates that, if not rejecting any true null hypothesis guarantees no incorrect decision, then controlling the probability of incorrectly rejecting at least one true null hypothesis (the so-called familywise error rate, or FWER) controls the rate of incorrect decision. Even though one would not want to assume monotonicity of efficacy in the MED setting, decision-making for MED does tend to follow a pre-specified path. Therefore, Hsu and Berger (1999) contains a principle, now called the Partitioning Principle, of how to formulate the null hypotheses so that FWER is controlled, and the resulting inference will automatically follow the desired decision path. The Partitioning Principle shows no multiplicity adjustment is needed in testing the null hypotheses along the path. This practice of not adjusting for multiplicity in testing along a decision path has now been applied to testing for efficacy in primary and secondary endpoints, a development Hsu and Berger (1999) did not anticipate in a setting for which FWER is an inadequate description of incorrect decision rates. We show that the marginal error rate of inference on the primary endpoint, together with the conditional

error rate of inference on the secondary endpoint (conditional on inferring efficacy in the primary), provide useful and directly interpretable assessment of incorrect decision rates of the current practice.

Email: hxu22@its.jnj.com

FAM2-1 T3

Optimal Statistical Decision for Gaussian Graphical Model Selection

Petr Koldanov, Alexander Koldanov, Valery Kalyagin, Panos Pardalos
NRU Higher School of Economics, Russia

Gaussian graphical model is a graph representation of dependence structure for Gaussian random vector (Edwards, D.2000, Lauritzen S.L.1996). It is recognized as powerful tool in different applied fields such as bioinformatics, error-control codes, speech language and information retrieval and others (Jordan M.I, 2004). Gaussian graphical model selection is statistical problem to identify Gaussian graphical model from sample of given size. Different approaches for Gaussian graphical model selection are suggested in the literature. One of them is based on considering the family of individual conditional independence tests (Dempster A P, 1972). Application of this approach leads to construction of variety of multiple testing statistical procedures for Gaussian graphical model selection (Drton M. Perlman M., 2007). Important characteristic for these procedures is its error rate for given sample size. In existing literature the great attention is paid to control of error rates for incorrect edge inclusion (Type I error). However in graphical model selection it is important to take into account error rates for incorrect edge exclusion too (Type II error). To handle this issue we consider graphical model selection problem in the framework of multiple decision theory. Quality of statistical procedures is measured by risk function with additive losses (Lehmann E.L., 1957). Additive losses allow to take into account both types of errors. We construct optimal unbiased tests of Neyman structure (Koldanov et al., 2017) for individual hypotheses and combine it to obtain a multiple decision statistical procedure. We show that obtained procedure is optimal in the sense that it minimizes linear combination of expected numbers of Type I and Type II errors in the class of unbiased multiple decision procedures. Detailed results of the talk are given in (Kalyagin et al. 2017).

References:

1. Edwards, D. Introduction to graphical modeling. Springer-Verlag New York, Inc., 2000.
2. Lauritzen S.L. Graphical models, Oxford University Press, N-Y, 1996.
3. Jordan M.I. Graphical Models, Statistical Science, v. 29 (2004), No.1, pp. 140-155.
4. Dempster A P Covariance selection. Biometrics, Vol. 28, (1972), pp. 157-175.
5. Drton M. Perlman M. Multiple Testing and Error Control in Gaussian Graphical Model selection, Statistical Science, Vol. 22 (2007), No. 3, pp.430-449.
6. Lehmann E.L. A theory of some multiple decision procedures 1, Annals of Mathematical Statistics, 28, 1_25 (1957).
7. V. A. Kalyagin, A. P. Koldanov, P. A. Koldanov, P. M. Pardalos Optimal statistical decision for Gaussian graphical model selection. (2017) arXiv:1701.02071
8. P. Koldanov, A. Koldanov, V.Kalyagin, P. Pardalos Uniformly most powerful unbiased test for conditional independence in Gaussian graphical model. Statistics & Probability Letters. 2017

Email: pkoldanov@hse.ru

FAM2-1 T4

Rank Verification for Exponential Families

Kenneth Hung, William Fithian
University of California, Berkeley, USA

Many statistical experiments involve comparing multiple population groups. For example, a public opinion poll may ask which of several political candidates commands the most support; a social scientific survey may report the most common of several responses to a question; or, a clinical trial may compare binary patient outcomes under several treatment conditions to determine the most effective treatment. Having observed the "winner" (largest observed response) in a noisy experiment, it is natural to ask whether that candidate, survey response, or treatment is actually the "best" (stochastically largest response). This article concerns the problem of rank verification-post hoc significance tests of whether the orderings discovered in the data reflect the population ranks. For exponential family models, we show under mild conditions that an unadjusted two-tailed pairwise test comparing the top two observations (i.e., comparing the "winner" to the "runner-up") is a valid test of whether the winner is truly the best. We extend our analysis to provide equally simple procedures to obtain lower confidence bounds on the gap between the winning population and the others, and to verify ranks beyond the first.

Email: kenhung@berkeley.edu

FAM2-2 T1

Use of Interval Estimations in Design and Evaluation of Multi-Regional Clinical Trials

Chin-Fu Hsiao, Chieh Chiang, H.M. James Hung
Institute of Population Health Sciences, National Health Research Institutes, Taiwan

Multi-regional clinical trials (MRCTs) have been promoted in recent years as a useful means of accelerating the development of new drugs and abridging their approval time. Global collaboration in MRCTs unites patients from several countries/regions around the world under the same protocol. The statistical properties of MRCTs have been widely discussed. However, when regional variability is taken into consideration, the assessment of efficacy response becomes much more complex. The current study represents an evaluation of the efficacy response for MRCTs based on Howe's, Cochran-Cox's, and Satterthwaite's interval estimations, which have been shown to have well-controlled type I error rates with heterogeneous regional variances. Corresponding sample size determination to achieve a desired power based on these interval estimations is also represented. Moreover, the consistency criteria suggested by the Japanese Ministry of Health, Labour and Welfare (MHLW) guidance to decide whether the overall results from the MRCT, via the proposed interval estimation, can be applied to a specific region or all regions are also derived. An example for three regions is used to illustrate the proposed method. Results of simulation studies are reported so that the proposed method can help determine the sample size and correctly evaluate the assurance probability of the consistency criteria.

Email: chinfu@nhri.org.tw

FAM2-2 T2

Multi-regional Biosimilarity Studies

Victoria Chang, Qi Xia
Boehringer-Ingelheim Pharmaceuticals Inc., USA

In biosimilarity drug development, the evidence of demonstrating biosimilarity is different from the evidence of approving a new drug. Biosimilarity is based on the totality of evidence of multiple steps of assessment. The fundamental steps consist of similarity in analytical and pharmacokinetic assessment. Some of the regional regulatory agencies may allow the sponsor to plan a phase III clinical trial with a single regional reference product instead of references by each region. Since the reference biological drugs marketed in different regions have not been demonstrated for biosimilarity, justification of using a reference product from regions other than the review region would require evidence of bridging between references marketed in different region. The bridging evidence needs to be established in analytical and pharmacokinetic assessment. In this paper, we reviewed the various setups and designs of the first four FDA approved biosimilar products and discussed the potential involvement of multiple comparisons that require type I error rate adjustment and power reduction. The impact may increases when more regional references are involved.

Email: changvick@gmail.com

FAM2-2 T3

MRCT design models and drop-min data analysis.

Fei Chen, K. K. Gordon Lan, Gang Li
Janssen R&D, USA

In recent years, developing pharmaceutical products via a multiregional clinical trial (MRCT) has become more popular. Many studies with proposals on design and evaluation of MRCTs under the assumption of a common treatment effect across regions have been reported in the literature. However, heterogeneity among regions causes concern that the fixed effects model for combining information may not be appropriate for MRCT. In this presentation, we will discuss: 1. The use of the fixed effect model, the continuous random effect model, and the discrete random effect model for the design and data analysis of MRCTs. Numerical examples will be provided to illustrate the fundamental differences among these three models. 2. Consistency and inconsistency: We will provide examples of inconsistency, and discuss the use of drop-the-min data analysis when the region with minimum treatment effect is excluded from the MRCT data analysis. We provide a solution first formulated within the fixed effects framework, and then extend it to discrete random effects model.

Email: FChen6@its.jnj.com

FAM2-3 T1

Comparing Several Variances with Control Using Sample Quasi Range

Rajvir Singh, Parminder Singh
Thapar University, India

In this research article, a class of step up test procedures for testing the homogeneity of scale parameters from k (k_1) normal populations with that of control population is proposed using sample quasi ranges. A recursive algorithm is used to compute the critical constants required for the implementation of proposed procedures. Simulation studies demonstrated that the proposed procedures are better than the existing competitors and are robust when outliers are present. The illustration of the proposed procedures is being done using a numerical data.

Email: rajvir.singh@thapar.edu

FAM2-3 T2

Revisiting "What's Wrong with Bonferroni Adjustments"

Andrew V. Frane
University of California, Los Angeles, USA

Ironically, the most frequently cited paper on the Bonferroni procedure is Thomas Perneger's "What's wrong with Bonferroni adjustments" (BMJ, 1998), which argued that the procedure should not be used. Perneger's paper has been cited thousands of times (typically by researchers seeking to justify their unadjusted statistical tests) and its influence has not waned over time. In fact, the paper has maintained a remarkably broad and enduring impact-garnering over 30 citations in 2016 alone, many of which appear in highly regarded scientific journals. Some of Perneger's arguments have been echoed even by authors who recognize the importance of addressing multiplicity in general. For instance, many proponents of multistep approaches to familywise error control have argued that the classical Bonferroni procedure is inherently overly conservative, especially when the tests are "planned" a priori, or when the tests have certain dependence structures, or when the "universal null hypothesis" is not of interest. Using both analytic and simulation-based methods, the current presentation critically evaluates the arguments in Perneger's often cited paper and makes a case for the appropriateness of Bonferroni adjustment in many common situations. Particular emphasis is placed on a unique and relatively unsung benefit of the classical Bonferroni procedure: control of the per-family error rate (i.e., the expected number of Type I errors per family). Additionally, the tradeoff between power and false discoveries is examined for established multiple-comparisons procedures (e.g., Bonferroni, Hochberg, Benjamini-Hochberg) under different combinations of parameters, sample sizes, and dependence structures.

Email: avfrane@gmail.com

FAM2-3 T3

The Reliability of Two Meta-Analysis Studies

Stan Young, Cheng You
CGStat, USA

Many regulatory decisions are based on meta-analysis of observational studies. There is a need to understand the reliability of meta-analysis studies. Our idea is to examine the reliability of the base studies used in two meta-analysis studies, one appearing in Lancet and the other in JAMA. Both of these studies examine the claimed causal effect of air quality on heart attacks. We count the number of outcomes, predictors, covariates and lags used in each base paper. Lags are of interest as the air quality yesterday might have a health effect today. Outcomes, predictors, and lags are used to estimate multiplicity. Covariates are used to estimate the number of possible models. Together they can be used to estimate the analysis search space available to the researcher. Altogether we examined 21 base papers. We find a median of 11,520 possible analyses with an interquartile range of 1,440 to 81,920. We conclude that the base papers do not support their claims due to a very large search space and that therefore the meta-analysis paper claims are not supported either. The benefit of our work is to inform regulatory bodies that previous regulations are not supported by papers using sound statistical analysis.

Email: genetree@bellsouth.net

FAM2-3 T4

Simultaneous Rank Tests for Pairwise Comparisons in Analysis of Covariance

Hossein Mansouri, Fangyuan Zhang,
Texas Tech University, USA

For the one-way analysis of variance, the method of simultaneous pairwise comparisons of treatment effects based on pairwise rankings of the samples provides a robust method of pairwise comparisons that controls the familywise error rate strongly. This is in comparison to the method of pairwise comparisons based on the overall ranking of all of the samples that is known to control the familywise error rate weakly. In this presentation, we will extend the method of simultaneous pairwise comparisons based on pairwise rankings of the samples to the analysis of covariance. Since the method is based on ranking after adjustment for covariates, it is based on large sample approximation theory. However, our simulation study indicates that the method has the robustness of validity property for small samples. This method also improves the control of the familywise error rate for the corresponding distribution-free pairwise comparisons when exact tables of the sampling distribution is not available and large sample approximation method is used.

Email: HOSSEIN.MANSOURI@ttu.edu

FAM2-3 T5

Adaptive Designs in Clinical Trials

Ramaiyan Elangovan
Annamalai Univarsity, India

An adaptive design is defined as a design that allows adaptations to trial and/or statistical procedures of the trial after its initiation without undermining the validity and integrity of the trial. In recent years, the use of adaptive design methods in clinical research and development based on accrued data has become very popular due to its flexibility and efficiency. Multiple comparison procedures using adaptive design methods in clinical trials has received much attention in recent literature and prevent the experimenter from declaring an effect when there is none. In this paper it is proposed to discuss the adaptive design methods in clinical trials using multiple comparison procedures. Numerical examples are substantiated through real data example concerning the development of Velcade, intended for multiple myeloma is also provided. Strategies for the use of adaptive design in clinical development of rare diseases using R software are also discussed.

Email: srelangovan@rediffmail.com

Presenters Index

Hobbs	Brian	ThPM1-3	15	65
Hsiao	Chin-Fu	FAM2-2	17	77
Hsiao	Sam	ThPM1-1	15	60
Hsu	Jason	WAM1-2	8	21
Hung	Kenneth	FAM2-1	17	77
Jaki	Thomas	ThPM1-2	15	62
Javanmard	Adel	WPM2-1	11	33
Jennison	Christopher	ThAM1-1	13	
Jiang	Lingling	ThAM1-2	13	49
Kamakura	Toshinari	FAM1-3	16	74
Katsevich	Eugene	WPM1-2	10	29
Kil	Siyoen	ThPM1-2	15	63
Klinglmueller	Florian	WAM2-1	9	22
		ThAM2-1	14	53
Koldanov	Petr	FAM2-1	17	76
Komiyama	Junpei	ThAM1-3	13	53
Lee	Michael	FAM1-2	16	71
Lei	Lihua	WAM2-2	9	23
		WPM3-1	12	46
Li	David	FAM1-2	16	72
Li	Huajiang	WAM2-3	9	24
Li	Huiling	WPM1-3	10	32
Liu	Yushi	ThPM1-1	15	61
Ma	Shujie	WPM1-1	10	28
Markovic	Jelena	FAM1-1	16	67
McCann	Melinda	WPM3-1	12	44
McTague	Jaclyn	ThPM1-1	15	61
Mehta	Cyrus	ThAM2-1	14	55
Meskaldji	Djalel-Eddine	ThAM2-2	14	57
Mütze	Tobias	WPM2-3	11	38
		ThAM1-1	13	48
Neumann	Andre	WPM3-1	12	40
Neuvial	Pierre	WAM2-3	9	26
Panigrahi	Snigdha	FAM1-1	16	68
Pecanka	Jakub	ThAM2-2	14	57
Pennello	Gene	WPM1-1	10	27
Posch	Martin	WAM2-1	9	22
		WPM2-3	11	39
Proschan	Michael	ThAM1-1	13	48
Qiu	Xing	FAM1-3	16	73
Rabinovich	Maxim	ThAM2-3	14	59
Ramdas	Aaditya	ThAM2-3	14	59
Robertson	David	ThAM2-1	14	54

Roquain	Etienne	WAM2-2	9	23
Rudra	Pratyaydipta	WPM3-1	12	43
Sackrowitz	Harold	WPM1-2	10	30
Saha	Saswati	FAM1-2	16	71
Shan	Guogen	ThAM1-2	13	50
Singh	Rajvir	FAM2-3	17	79
Sirotko-Sibirskaya	Natalia	WPM3-1	12	42
Sivaganesan	Siva	WPM1-1	10	28
Solari	Aldo	WPM1-2	10	30
Song	Yanglei	WPM2-1	11	35
Sugiyama	Mahito	ThAM1-3	13	52
Su	Weijie	ThAM 2-3	14	58
Umezu	Yuta	WPM3-1	12	46
Tamhane	Ajit	WPM1-3	10	31
Takeuchi	Ichiro	WPM2-2	11	36
Tayob	Nabihah	ThPM1-3	15	65
Tang	Szu-Yu	WPM1-1	10	27
Tsuda	Koji	WPM2-2	11	35
Vickersta	Victoria	FAM1-2	16	70
Vlachos	Pantelis	ThPM1-1	15	60
Wang	Bushi	WAM2-1	9	22
		WPM1-3	10	31
Wang	Jingshu	FAM1-3	16	72
Webb	Geoff	WPM2-2	11	37
Weinstein	Asaf	FAM1-1	16	69
Wittes	Janet	ThAM1-1	13	47
Wolf	Michael	WAM2-2	9	24
Xi	Dong	WAM2-1	9	22
		ThAM2-2	14	56
Xu	Haiyan	WAM2-1	9	22
		FAM2-1	17	75
Xu	Ningning	WPM3-1	12	45
Yagi	Ayaka	WPM3-1	12	41
Yang	Fanny	WPM2-3	11	40
		WPM3-1	12	45
Young	Stan	FAM2-3	17	80
Mansouri	Hossein	FAM2-3	17	80
Zhang	Zhiwei	ThPM1-2	15	64
Zhang	Qi	FAM1-3	16	73
Zhu	Jian	WPM2-3	11	37