International Chinese Statistical Association

Applied Statistics Symposium

2017

CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS

June 25 - 28, 2017
Hilton Chicago
Chicago, Illinois, USA

Organized by
International Chinese Statistical Association
## Contents

<table>
<thead>
<tr>
<th>Welcome</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conference Information</td>
<td>2</td>
</tr>
<tr>
<td>ICSA Officers and Committees</td>
<td>2</td>
</tr>
<tr>
<td>Conference Committees</td>
<td>5</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>8</td>
</tr>
<tr>
<td>Travel and Transportation</td>
<td>13</td>
</tr>
<tr>
<td>Conference Venue Information</td>
<td>14</td>
</tr>
<tr>
<td>Program Overview</td>
<td>19</td>
</tr>
<tr>
<td>Keynote Lectures</td>
<td>21</td>
</tr>
<tr>
<td>Student Paper Awards</td>
<td>26</td>
</tr>
<tr>
<td>Short Courses</td>
<td>27</td>
</tr>
<tr>
<td>ICSA 30-Year Celebration Panel Session</td>
<td>37</td>
</tr>
<tr>
<td>Social Events</td>
<td>38</td>
</tr>
<tr>
<td>ICSA 2018 in New Brunswick New Jersey</td>
<td>39</td>
</tr>
<tr>
<td>2017 ICSA and ASA Fall Annual Meeting</td>
<td>40</td>
</tr>
<tr>
<td>Scientific Program</td>
<td>41</td>
</tr>
<tr>
<td>Sunday, June 25. 1:00 PM - 2:40 PM</td>
<td>41</td>
</tr>
<tr>
<td>Sunday, June 25. 3:15 PM - 4:55 PM</td>
<td>43</td>
</tr>
<tr>
<td>Monday, June 26. 8:30AM - 9:45 AM</td>
<td>44</td>
</tr>
<tr>
<td>Monday, June 26. 10:00 AM - 11:40 AM</td>
<td>44</td>
</tr>
<tr>
<td>Monday, June 26. 1:00 PM - 2:40 PM</td>
<td>48</td>
</tr>
<tr>
<td>Monday, June 26. 2:50 PM - 4:30 PM</td>
<td>51</td>
</tr>
<tr>
<td>Monday, June 26. 4:45PM - 5:45 PM</td>
<td>55</td>
</tr>
<tr>
<td>Tuesday, June 27. 8:00 AM - 10:10 AM</td>
<td>55</td>
</tr>
<tr>
<td>Tuesday, June 27. 10:30 AM - 12:10 PM</td>
<td>55</td>
</tr>
<tr>
<td>Tuesday, June 27. 1:30 PM - 3:10 PM</td>
<td>59</td>
</tr>
<tr>
<td>Tuesday, June 27. 3:30 PM - 5:10 PM</td>
<td>62</td>
</tr>
<tr>
<td>Wednesday, June 28. 8:30 AM - 10:10 AM</td>
<td>66</td>
</tr>
<tr>
<td>Wednesday, June 28. 10:30 AM - 12:10 PM</td>
<td>69</td>
</tr>
<tr>
<td>Abstracts</td>
<td>73</td>
</tr>
<tr>
<td>Session 1: Recent progress in high dimensional data analysis</td>
<td>73</td>
</tr>
<tr>
<td>Session 2: Spatial and Temporal Modeling for Environmental Data</td>
<td>73</td>
</tr>
<tr>
<td>Session 3: Recent advances in statistical methods for brain connectome networks</td>
<td>74</td>
</tr>
<tr>
<td>Session 4: New Bayesian methods in applied statistics</td>
<td>75</td>
</tr>
<tr>
<td>Session 5: Advance in Statistical Methods for Large and Complex Data</td>
<td>76</td>
</tr>
<tr>
<td>Session 6: Recent Advances for Handling Composite Endpoints in Clinical Trials - Statistical Methodologies and Applications</td>
<td>76</td>
</tr>
<tr>
<td>Session 7: New methods development in Genetic and genomic data analysis</td>
<td>77</td>
</tr>
<tr>
<td>Session 8: Recent developments in multiple testing</td>
<td>78</td>
</tr>
<tr>
<td>Session 9: Novel Analysis of Big Data arising from mHealth and Sensor Technology</td>
<td>78</td>
</tr>
<tr>
<td>Session 10: Recent advances in methods for EHR based research</td>
<td>79</td>
</tr>
<tr>
<td>Session 11: Recent Advances in High-dimensional Inference</td>
<td>80</td>
</tr>
<tr>
<td>Session 12: Survival Analysis and Genetics</td>
<td>81</td>
</tr>
<tr>
<td>Session 13: Design and Analysis of Dose Finding Clinical Trials for Combination Therapy</td>
<td>81</td>
</tr>
<tr>
<td>Session 14: Understanding the Microbiome Complexity – Genetics and Networks</td>
<td>82</td>
</tr>
</tbody>
</table>

---

iii
Session 71: General Developments in Nonclinical Statistics .................................................. 122
Session 72: Topics on statistical modeling and inference (I) .................................................. 123
Session 74: Statistical and Computational Challenges for Single-cell Sequencing ................. 123
Session 75: Several advanced topics in statistical metagenomics ........................................ 124
Session 76: Recent Advancements in Personalized Medicine ............................................... 125
Session 77: Statistical Tests for Data Measured on Taxonomic Trees in Microbiome Studies .... 125
Session 78: Complex and Large Data Analysis .................................................................... 126
Session 79: Inference for Big and Complex Data ................................................................. 127
Session 80: Quantile regression and inference for high-dimensional problems ..................... 128
Session 81: Recent Advances on Statistical Modeling for Cancer Genomics Data ............... 128
Session 82: Statistical considerations in Multi-regional trials ............................................. 129
Session 83: Combining Endpoints in Clinical Trials ........................................................... 130
Session 84: Biomarker development: Ideas and practicalities ............................................. 131
Session 85: Statistical Analysis for Non-normal Data .......................................................... 131
Session 86: Statistical methods in oncology trials ............................................................... 132
Session 87: Mixed-effects models in genomics, health and psychology ............................... 133
Session 88: Making Sense of Omics Data: Recent Advances in Statistical and Computational Methods .......................................................... 133
Session 89: Statistical Applications in Population Health Sciences .................................... 134
Session 90: New Advances on Nonparametric Methods for High-dimensional Data ............. 135
Session 91: Student session (II) ......................................................................................... 136
Session 92: New Advances in Modeling of Functional Data .................................................. 137
Session 93: Statistical and Computational Approaches for High-throughput Sequencing Data Analysis .................................................................................. 138
Session 94: Non-clinical and biosimilar .............................................................................. 138
Session 95: Statistical Methods for Modelling Data Complexity in Genomics Studies ......... 139
Session 96: Recent advances in early phase oncology trials using drug combinations .......... 140
Session 97: Advanced Statistical Methods for Complicated High Dimensional Data ............. 140
Session 98: Advances in Low-rank Modeling and Its Estimation ......................................... 141
Session 99: Recent Advancement and Applications in Multiplicity Adjustments .................. 142
Session 100: Design and analysis of cancer immunotherapy trials ....................................... 142
Session 101: Design and practical considerations for dose selection trials ........................... 143
Session 102: Novel Statistical Methods for Analysis of Complex Biological Data ................ 144
Session 103: TBD ............................................................................................................. 144
Session 104: Analysis of High-Dimensional Data and Applications ..................................... 145
Session 105: Recent advances in clinical trial design and analysis with time to event endpoints .......................................................... 146
Session 106: Methods and applications in clustered and high dimensional data .................. 146
Session 107: Innovative clinical trial designs ..................................................................... 147
Session 108: Topics on Applied Statistics (I) ..................................................................... 148
Session 109: Topics on structured models .......................................................................... 149
Session 110: Advances in change-point analysis ................................................................. 150
Session 111: The challenges of non-constant hazard ratio: delayed treatment effect, treatment dilution and treatment crossover .......................................................... 150
Session 112: Adaptive Design .......................................................................................... 151
Session 113: Recent Development and Applications of Statistical Methodology in Rare Diseases .......................................................... 152
Session 114: Statistical Genetics & Genomics (or Computational Biology) ....................... 153
Session 115: Recent Advances in Dose Finding studies ....................................................... 153
Session 116: Partial Identification: Seeking reasoned evidence from difficult data ............. 154
Session 117: High dimensional multivariate analysis and its applications ......................... 154
Session 118: Recent Developments in Graphical Models and Network Analysis ................ 155
Session 119: Novel statistical methods for genetic data analysis ........................................ 156
Session 120: Complex Data Analysis; Methodologies and Applications .............................. 157
Session 121: Multiplicity in Clinical Trials ......................................................................... 157
Session 122: high dimensional modeling in medicine ......................................................... 158
Session 123: Utilizing Real World Evidence in Regulatory Decision: Statistical Considerations and Beyond .......................................................... 159
Session 124: Uncertainty, Effect Size and Bias in Statistical Evidence ................................. 159
Session 125: Recent experience in accelerated approval in Oncology Immunotherapy from Industrial and FDA perspective .................................................................................. 160
Welcome to the 2017 International Chinese Statistical Association (ICSA) Applied Statistical Symposium! It is the 26th ICSA annual symposium. The theme of this symposium is Statistics for a new generation: challenges and opportunities, in recognition of the advent of a new generation of statisticians. The executive and organizing committees have been working diligently to put together a comprehensive and exciting program including 11 short courses, four keynote addresses, 160 scientific sessions, and two student paper sessions. This year marks ICSA 30-year anniversary and we will celebrate with a designated panel session and exciting social events.

Our scientific program includes keynote addresses from renowned statisticians Dr. Xiaoli Meng (Harvard University), Dr. Rod Little (University of Michigan) and Dr. Ram Tiwari (FDA), and Barry Nussbaum (ASA President), invited, topic contributed, and contributed talks, presenting cutting edge advances of statistical theory, methods and application in all areas. Social events in this 2017 ICSA Symposium include mixer (Monday, June 26 evening) and banquet (Tuesday, June 27). To celebrate 30-year anniversary, an entertainment program will be offered at the banquet. To the end of our theme, we are offering a full career service including onsite interview from leading pharmaceutical companies and recruiting firm.

With your enthusiastic participation, this symposium attracts more than 800 statisticians in academia, government, and industry from all over the world. We are confident that it offers great opportunities for learning, sharing, networking and relaxing.

June is the most agreeable time for the city of Chicago, which is directly accessible from most cities worldwide. Downtown Chicago provides top notch dining, shopping and lodging experiences. In addition, it offers world-class attractions, including the Sky Deck, the Millennium Park, the Navy Pier, and the Shedd Aquarium and numerous museums, all within walk distance from Hilton Chicago.

Thank you for coming to the 2017 ICSA Applied Statistics Symposium!

Lanju Zhang
on behalf of 2017 Applied Statistics Symposium Executive and Organizing committees
ICSA OFFICERS and COMMITTEES (January-December 2017)

EXECUTIVES

President: T. Tony Cai (tcai@wharton.upenn.edu)
Past President: Mei-Ling Ting Lee (mltlee@umd.edu)
President-elect: Aiyi Liu (liua@mail.nih.gov)
Executive Director: Gang Li (2017-2019, gli@its.jnj.com)
ICSA Treasurer: Hongliang Shi (2016-2018, hongliangshi15@gmail.com)

BOARD of DIRECTORS

Qu, Annie (2017, Biometrics Section Representative, anniequ@illinois.edu)
Shuangge Ma (2015-2017, shuangge.ma@yale.edu)
Dejun Tang (2015-2017, dejun.tang@sandoz.com)
Lilly Yue (2015-2017, lilly.yue@fda.hhs.gov)
Ying Yuan (2015-2017, yyuan@mdanderson.org)
Tian Zheng (2015-2017, tzheng@stat.columbia.edu)
Alan Chiang (2016-2018, chiangay@lilly.com)
Yanyuan Ma (2016-2018, yanyuannma@stat.sc.edu)
Lu Tian (2016-2018, lutian@stanford.edu)
Bo Yang (2016-2018, byang2233@gmail.com)
Hao Helen Zhang (2016-2018, hzhang@math.arizona.edu)
Jie Chen (2017-2019, jiechen0713@gmail.com)
Jianhua Hu (2017-2019, jhu@mdanderson.org)
Bo Huang (2017-2019, bo.huang@pfizer.com)
Chuang-Yu Huang (2017-2019, cyhuang@jhu.edu)
Wang, Sijian (2017-2019, swang@biostat.wisc.edu)

STANDING COMMITTEES

Program Committee:

Ming Yuan (Chair, 2017, UW-Madison, ming.mingyuan@gmail.com),
Yichuan Zhao (2015-2017, yichuan@gsu.edu),
Ying Zhang (2015-2017, yz73@iu.edu),
Ming-Hui Chen (2015-2017, ming-hui.chen@uconn.edu),
Yi Li (2016-2018, yili@umich.edu),
Lanju Zhang (2016-2018, Lanju.zhang@abbvie.com),
Hongmei Jiang (2016-2018, hongmei@northwestern.edu),
Wenguang Sun (2017-2019, USC, wenguans@marshall.usc.edu),
Dan Yang (dyang@stat.rutgers.edu),
Hong Tian (2017-2019, HTian@its.jnj.com).

Finance Committee:

Hongliang Shi (Chair, 2016-2018, hongliangshi15@gmail.com),
Linda Yau (2016-2018, yau.linda@gene.com),
ICSA Officers and Committees

Rochelle Fu (2015-2017, fur@ohsu.edu).

Nomination and Election Committee:

Tianxi Cai (Chair, 2017, Harvard, tcai@hsph.harvard.edu),  
Haoda Fu (2015-2017, fu_haoda@lilly.com),  
Yuan Ji (2015-2017, YJi@health.bsd.uchicago.edu),  
Li-Shan Huang (2015-2017, lhuang@stat.nthu.edu.tw),  
Xuewen Lu (2015-2017, xlu@ucalgary.ca),  
Joanna Shih (2016-2018, jshih@mail.nih.gov),  
Tsai-Hung Fan (2016-2018, thfanncu@gmail.com),  
Hongzhe Li (2017-2019, hongzhe@mail.med.upenn.edu),  
Niansheng Tang (2017-2019, nstang@ynu.edu.cn)

Publication Committee:

Yichuan Zhao (2017-2019, yichuan@gsu.edu),  
Chunming Zhang (2015-2017, cmzhang@stat.wisc.edu),  
Kelly Zou (2015-2017, Kelly.Zou@pfizer.com),  
Xin He (Editor of Bulletin),  
Mei-Cheng Wang (Co-Editor of SIB, mcwang@jhu.edu),  
Hongzhe Li (Co-Editor of SIB, hongzhe@mail.med.upenn.edu),  
Hsin-Cheng Huang (Co-Editor of S. Sinica),  
Ruey Tsay (Co-Editor of S. Sinica),  
Zhiliang Ying (Co-Editor of S. Sinica),  
Gang Li (Ex-Officio),  
Jiahua Chen (Editor of ICSA book series),  
Din Chen (Co-Editor of ICSA book series)

Membership Committee:

Tony Jianguo Sun (Chair, 2017, sunj@missouri.edu),  
Bo Fu (2016-2018, bo.fu@abbvie.com),  
Zhigen Zhao (2015-2017, zhaozhg@temple.edu),  
Lanju Zhang (2015-2017, lanju.zhang@abbvie.com),  
Jiajuan Liang (2016-2018, JLiang@newhaven.edu),  
Samuel Wu (2016-2018, sw45@ufl.edu),  
Pengsheng Ji (2017-2019, psji@uga.edu),  
Yong Chen (2017-2019, ychen123@upenn.edu),  
Rui Feng (2017-2019, ruifeng@mail.med.upenn.edu).

Awards Committee:

Linda Zhao (Chair, 2017-2019, UPenn, lzhao@wharton.upenn.edu),  
Shu-Yen Ho (2015-2017, shu-yen.y.ho@gsk.com),  
Ming Yen Cheng (2016-2018, cheng@math.ntu.edu.tw),  
Tsung-Chi Cheng (2016-2018, chengt@nccu.edu.tw),  
Yazhen Wang (2017-2019, yzwang@stat.wisc.edu),  
Zongming Ma (2017-2019, zongming@wharton.upenn.edu),
Yichuan Zhao (2017-2019, yichuan@gsu.edu),
Song-Xi Chen (2017-2019, songchen@iastate.edu),
Harrison Zhou (2017-2019, huibin.zhou@yale.edu).

**IT Committee:**
Chengsheng Jiang (Chair, 2017, chengsheng.jiang@gmail.com),
Lixin (Simon) Gao (gaolixin@yahoo.com),
Don Sun (dsun48@gmail.com),
Ruth Whitworth (rewhitworth@georgiasouthern.edu),
Hongtu Zhu (htzhu@email.unc.edu),
Gang Li (Ex-Officio)

**Archive Committee:**
Lili Yu (Chair, 2017, lyu@georgiasouthern.edu), Smiley Cheng, Shein-Chung Chow, Nancy Lo

**ICSA Representative to JSM Program Committee:**
Yi Li (2017, yili@umich.edu), Wenguang Sun (2018, wenguans@marshall.usc.edu).

**AD HOC COMMITTEES**

**2017 Applied Statistics Symposium Committee:** Lanju Zhang, Chair of the Executive Committee (lanju.zhang@abbvie.com) and Hongmei Jiang, Chair, the Local Committee.

**2017 Special Lectures Committee:** Qiman Shao (Chair, CUHK, qmshao@sta.cuhk.edu.hk), Yazhen Wang (UW-Madison, yzwang@stat.wisc.edu), Yin Xia (UNC/Fudan, xiayin@email.unc.edu)

**2017 Fund-Raising Committee:** Wei Shen (Chair, Eli Lilly, shen@lilly.com), Xiaoyu (Cassie) Dong (Amgen, xdong01@amgen.com), Chunming Mark Li (Pfizer, Chunming.M.Li@pfizer.com),

**2017 JSM Local Committee:** Chenguang Wang (Chair, cwang68@jhmi.edu)

**Investment Ad Hoc Committee:** Linda Yau (Chair, 2015, yau.linda@gene.com), Gang Li (Ex-Officio, gli@its.jnj.com), Fang Chen (FangK.Chen@sas.com), Hongliang Shi (hongliangshi15@gmail.com), Ming-Dauh Wang (wang_ming-dauh@lilly.com), Rochelle Fu (fur@ohsu.edu), Mingxiu Hu (Consultant, Mingxiu.Hu@takeda.com)

**CHAPTERS**

**ICSA-Canada Chapter:** Changbao Wu (Chair, cbwu@uwaterloo.ca), Wendy Lou (Treasurer/Secretary, wendy.lou@utoronto.ca)

**ICSA - New England Chapter:** Huyuan Yang (Chair, Huyuan.Yang@takeda.com)

**ICSA – Midwest Chapter:** Li Wang (li.wang@abbvie.com)

**ICSA Shanghai Committee (Term 10/2013 to 10/2016):** Dejun Tang (Chair, Novartis, China)

**Office of ICSA**
Lili Yu, Congjian Liu, Ruth Whitworth, Karl Peace,
Jiann-Ping Hsu College of Public Health, Georgia Southern University
Conference Committees

Executive Committee

Lanju Zhang  Executive Committee Chair, AbbVie Inc
Xuming He  Program Committee Chair, University of Michigan
Yuan Ji  Program Book Chair, University of Chicago
Hongmei Jiang  Local Committee Chair, Northwestern University
Gang Li  Strategic Advisor, Janssen Research and Development
Xuan Liu  Treasurer, AbbVie Inc
Wei Shen  Strategic Advisor, Eli Lilly and Company
Bo Yang  Fund Raising Chair, Vertex Pharmaceuticals
Lingsong Zhang  Student Paper Award Chair, Purdue University
Jun Zhao  Short Course Chair, AbbVie Inc

Local Committee

Hongmei Jiang  Chair, Northwestern University
Jun Zhao  Co-Chair, AbbVie Inc
Lanju Zhang  Co-Chair, AbbVie Inc
Jie Yang  University of Illinois at Chicago
Jingjing Gao  AbbVie Inc
Jiping Wang  Northwestern University
Juan Hu  DePaul University
Lei Liu  Northwestern University
Lihui Zhao  Northwestern University
Min Yang  University of Illinois at Chicago
Xin Wang  AbbVie Inc

Program Committee

Xuming He  Chair, University of Michigan
Haoda Fu  Co-Chair, Eli Lilly and Company
Lin Chen  University of Chicago
Yi-Hau Chen  Institute of Statistical Science Academia Sinica, Taiwan
Ken Cheung  Columbia University
Yuehua Cui  Michigan State University
Jianhua Hu  MD Anderson Cancer Center
Hongmei Jiang  Northwestern University
Mei-Ling Ting Lee  University of Maryland
Hongzhe Li  University of Pennsylvania
Lexin Li  Berkeley
Ruoshua Li  University of Texas
Yehua Li  Iowa State University
Yi Liu  Takeda
Tim O’Brien  Loyola University
Conference Committees

Jane Qian  AbbVie Inc
Hui Quan  Sanofi
Jianxin Shi  NCI
Yang Song  Merck
Dejun Tang  Novartis China
Hong Tian  Janssen Research and Development
Yi Tsong  FDA
Xinyi Xu  Ohio State University
Lingsong Zhang  Purdue University
Yanqiong Zhang  Agios Pharmaceuticals, Inc.
Dave Zhao  University of Illinois at Urbana-Champaign

Program Book Committee

Yuan Ji  Chair, University of Chicago
Shengjie Yang  NorthShore University HealthSystem
Jiaying Lyu  Fudan University

Student Paper Award Committee

Lingsong Zhang  Chair, Purdue University
Anindya Bhadra  Purdue University
Bo Li  UIUC
Ruoshan Li  The University of Texas School of Public Health
Huiyan Sang  Texas A&M University
Cheng Yong Tang  Temple University
Jin Zhou  Arizona State University

Short Course Committee

Jun Zhao  Chair, AbbVie Inc
Yuan Ji  University of Chicago
Hongmei Jiang  Northwestern University
Faming Liang  University of Florida
Xin Tian  NIH
Yijie Zhou  AbbVie Inc

Fund Raising Committee

Bo Yang  Chair, Vertex Pharmaceuticals
Alan Chiang  Eli Lilly and Company
Sammi Tang  Vertex Pharmaceuticals
Yijie Zhou  AbbVie Inc
# Conference Committees

## Webmaster

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shengjie Yang</td>
<td>NorthShore University HealthSystem (Conference webpage)</td>
</tr>
<tr>
<td>Jiping Wang</td>
<td>Northwestern University (Conference webpage)</td>
</tr>
<tr>
<td>Chengsheng Jiang</td>
<td>University of Maryland, College Park (ICSA webpage)</td>
</tr>
<tr>
<td>Michael Zhang</td>
<td>University of Minnesota (UIC Housing Reservation page)</td>
</tr>
</tbody>
</table>

## Thirty-year Anniversary Celebration Committee

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mei-Ling Ting Lee</td>
<td>Chair, University of Maryland</td>
</tr>
<tr>
<td>Tony Cai</td>
<td>University of Pennsylvania</td>
</tr>
<tr>
<td>Jeng-Min Chiou</td>
<td>Institute of Statistics Sciences, Academia Sinica, Taiwan</td>
</tr>
<tr>
<td>Ruey Tsay</td>
<td>University of Chicago</td>
</tr>
</tbody>
</table>
Acknowledgements

Symposium Sponsors
The 2017 ICSA Applied Statistics Symposium is supported by a financial contribution of the following sponsors:

Gold Sponsors:

Gold Sponsors:

General Sponsors:

General Sponsors:

The organizing committees greatly appreciate the support of the above sponsors.
EVERYDAY LIFE AT ABBVIE IS FULL OF PACE, PURPOSE AND PASSION.

AbbVie is a global, research-based biopharmaceutical company whose mission is to use its unique approach to innovation to develop advanced therapies that address some of the world’s most complex and serious diseases. AbbVie employs more than 28,000 people worldwide and markets medicine in more than 170 countries. As part of Data and Statistical Sciences, the Pipeline Statistics and Programming organization provides expertise in biostatistics, bioinformatics, and statistical programming in support of AbbVie’s pharmaceutical research and development.

To learn more, or to find opportunities to join our team of dedicated professionals, visit www.abbvie.com/careers.

PEOPLE. PASSION. POSSIBILITIES.
Boehringer Ingelheim is hiring Biostatisticians!!

At Boehringer Ingelheim our vision "Value through Innovation" shows our innovative, sustainable and successful entrepreneurial approach. To learn more about our pipeline and career opportunities, visit our website at...

https://careers.boehringer-ingelheim.com

Senior Biostatistician or Principal Biostatistician
You will perform duties of a Trial Statistician to support complex clinical trials within national or international development projects or for marketed products as required. Assure well-designed clinical trials. Provide statistical expertise necessary to design, analyze, interpret and communicate the results of complex clinical trials. Provide statistical input for publications on clinical trials. Either support Project Statisticians on complex projects or act as a Project Statistician for early projects, backup projects, or projects with established BI experience. Act as a team leader for a complex project or mega-trial.

https://boehringer.taleo.net/careersection/global+template+career+section+28external29/jobdetail.ftl?job=171780&lang=en&sns_id=mailto#.WNUaLDE3TPw.mailto

Senior Principal Biostatistician
You will perform duties of a Project or Trial Statistician to support high-profile international clinical development projects/trials or marketed products as required. Provide statistical input into the creation of clinical development plans and/or marketing and publication strategies. Assure well-designed clinical trials. Take on the key statistical responsibility in the planning and preparation of regulatory submissions. Represent the company as required before regulatory agencies and other organizations in defense of statistical methodologies and data analyses related to our products. Serve on internal Data Monitoring Committees (DMCs) or represent the company on external DMCs.

https://boehringer.taleo.net/careersection/global+template+career+section+28external29/jobdetail.ftl?job=1613322&lang=en&sns_id=mailto#.WNUa0IJrIQU.mailto

Distinguished Biostatistician, Oncology
You will fulfill the role of a Therapeutic Area Statistician participate in therapeutic area of Oncology and to serve as a Project Statistician for a high-profile project. To pursue statistical methodological research. To ensure quality and consistency of statistical contribution to clinical drug development on. Act as an internal statistical consultant on critical statistical issues in clinical drug development and regulatory science.

For 30 years Gilead Sciences has been revolutionizing healthcare by bringing urgently needed medicines to millions of patients in the areas of HIV/AIDS, liver diseases, hematology and oncology, inflammation and respiratory diseases and cardiovascular conditions.

As a result of continued growth driven by Gilead’s investment in research & development we are looking to invite those with a genuine commitment to patient care to play a critical role in helping to bring important new therapies to those with life-threatening illnesses.

Gilead’s unique operational environment allows exceptionally talented professionals gain rapid skills and knowledge growth whilst supporting Gilead’s patient driven mission.

Statisticians at Gilead provide statistical support on clinical trial design and drug development strategy, author statistical analysis plans, and support publications and scientific presentations. They are given the opportunity to work in an agile and collaborative environment and to advance innovative therapeutics by providing statistical foundation for drug development and clinical data excellence.

To find out more, connect with us today.

www.gilead.com/careers
Introduction to The Lotus Group

The Lotus Group is a professional recruiting firm which specializes in the pharmaceutical industry, specifically in statistician, statistical programming, data management and data science positions. We pride ourselves in providing excellent customer service and building solid relationships with our client companies so that we can provide the best opportunities for our candidates. As a result of our industry expertise and long-lasting relationships, we partner with the best pharmaceutical and biotech companies in the world and are able to recruit top quality talent for our clients. The Lotus Group team members are located in New Jersey, Boston, California, Pennsylvania and Chicago. We are effectively able to connect locally with candidates in strong pharmaceutical markets. We also regularly attend statistical and industry-related conferences, so we stay well-informed and well-connected.

We are hosting a career session at the 2017 Symposium in Chicago on Monday 6/26 at 2:50pm CT. We welcome everyone to join us and chat with our distinguished and influential panel speakers:

Ivan Chan, VP of Pipeline Statistics and Programming at AbbVie

Chito Hernandez, VP of Biometrics at Biomarin

Stacy Lindborg, VP of Biometrics at Biogen

Naitee Ting, Director, Statistics Technical, Boehringer Ingelheim Pharmaceuticals

Lei Wang, Managing Partner, The Lotus Group

The topic of our discussion is:

What Makes a Well Rounded Statistician in the Pharmaceutical Industry?

To be an influential biostatistician in the pharmaceutical industry requires a lot more than expertise in statistics. We are thrilled to have five industry leaders in biostatistics to share their vision and insights for what makes a successful statistician through the eyes of their own career experiences. Topics to be covered will include, for example: what are additional critical skills required to be an effective and visible statistician, tips on how to build a strong professional network, specific suggestions on how to grow in the industry from a junior individual contributor to a well-rounded statistician and leader.
**Travel and Transportation**

**Information**

Chicago is in Central Time Zone. For your convenience of scheduling the travel, the last talk on June 28, 2017 (Wednesday) will conclude at noon.

**Accommodation**

The 2017 Applied Statistics Symposium by the International Chinese Statistical Association (ICSA) will be held on June 25-28, 2017 at Hilton Chicago. The address is 720 S Michigan Ave, Chicago, IL 60605 USA

The Hilton hotel has blocked guest rooms at a discounted rate for attendees. Economy lodging is provided for attendees at University of Illinois at Chicago. Please go to registration page for more details and make a reservation.

**Airports**

- O'Hare International Airport
- Midway International Airport
Hilton Chicago
Hilton Chicago
Conference Venue Information

EIGHTH FLOOR
- Lake Ontario
- Lake Michigan
- Lake Huron
- Lake Erie

FIFTH FLOOR
- Conference Rooms 5A-5J

FOURTH FLOOR
- Conference Rooms 4A-4R
- McCormick Boardroom
- Pullman Boardroom

THIRD FLOOR
- Waldorf
- Astoria
- Williford
- Marquette
- Joliet
- Private Dining Rooms 1-7

SECOND FLOOR
- Grand Ballroom
- Grand Ballroom Foyer
- International Ballroom
- International Foyer
- Normandie Lounge
- Boulevard Room
- Boulevard Foyer

LOBBY LEVEL
- Continental Ballroom
- Continental Foyer
- Grand Tradition
- 8th Street South Registration
- 8th Street North Registration

LOWER LEVEL
- Salon A
- Salon B
- Salon C
- Salon D
- Mobley Room
- Registration

ROOM RELATIONSHIPS
## Program Overview

### Sunday June 25, 2017

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM - 6:00 PM</td>
<td>Waldorf Room</td>
<td>Registration</td>
</tr>
<tr>
<td>7:00 AM - 8:45 AM</td>
<td>Waldorf Room</td>
<td>Breakfast</td>
</tr>
<tr>
<td>9:45 AM – 10:15 AM</td>
<td></td>
<td>Break</td>
</tr>
<tr>
<td>8:00 AM - 5:00 PM</td>
<td>5A</td>
<td><strong>Short course</strong>: Quantitative Sciences for Safety Monitoring during Clinical Development</td>
</tr>
<tr>
<td>8:00 AM - 12:00 PM</td>
<td>5B</td>
<td><strong>Short course</strong>: Multi-Regional clinical trials (including textbook)</td>
</tr>
<tr>
<td>8:00 AM - 12:00 PM</td>
<td>5C</td>
<td><strong>Short course</strong>: Methods of biomarker and subgroup identification for personalized medicine</td>
</tr>
<tr>
<td>8:00 AM - 12:00 PM</td>
<td>5D</td>
<td><strong>Short course</strong>: Bayesian Adaptive Clinical Trials in the 21st Century</td>
</tr>
<tr>
<td>8:00 AM - 12:00 PM</td>
<td>5E</td>
<td><strong>Short course</strong>: Statistical methods and software for multivariate meta-analysis</td>
</tr>
<tr>
<td>8:00 AM - 12:00 PM</td>
<td>5F</td>
<td><strong>Short course</strong>: Clinical Trials for Time-to-Event Outcomes: Current Practice and New Developments</td>
</tr>
<tr>
<td>12:00 PM - 1:00 PM</td>
<td>Waldorf Room</td>
<td>Lunch for Registered Full-Day Short Course Attendees</td>
</tr>
<tr>
<td>1:00 PM - 5:00 PM</td>
<td>5B</td>
<td><strong>Short Course</strong>: Integrative Multivariate Statistical Learning: Recent Developments and Case Studies</td>
</tr>
<tr>
<td>1:00 PM - 5:00 PM</td>
<td>5C</td>
<td><strong>Short Course</strong>: Phase II Clinical Development</td>
</tr>
<tr>
<td>1:00 PM - 5:00 PM</td>
<td>5D</td>
<td><strong>Short Course</strong>: Principles of Multiple Comparisons, with Applications</td>
</tr>
<tr>
<td>1:00 PM - 5:00 PM</td>
<td>5E</td>
<td><strong>Short Course</strong>: Bayesian Adaptive Phase I Oncology Trials: Methodology and Implementation</td>
</tr>
<tr>
<td>1:00 PM - 5:00 PM</td>
<td>5F</td>
<td><strong>Short Course</strong>: Design and Analysis of Real-word Clinical Studies Using SAS and R</td>
</tr>
<tr>
<td>1:00 PM - 5:00 PM</td>
<td>5G</td>
<td><strong>Short Course</strong>: Experiences and Case Studies in Adaptive Clinical Trial Design</td>
</tr>
<tr>
<td>2:45 PM - 3:15 PM</td>
<td></td>
<td>Break</td>
</tr>
<tr>
<td>1:00 PM - 2:40 PM</td>
<td>See program</td>
<td><strong>Parallel Session</strong></td>
</tr>
<tr>
<td>3:15 PM - 4:55 PM</td>
<td>See program</td>
<td><strong>Parallel Session</strong></td>
</tr>
<tr>
<td>4:30 PM - 6:00 PM</td>
<td>Astoria</td>
<td><strong>ICSA Board Meeting (Invitation Only)</strong></td>
</tr>
<tr>
<td>6:30 PM - 8:30 PM</td>
<td></td>
<td><strong>Off site Board Dinner (Invitation Only)</strong></td>
</tr>
</tbody>
</table>
# Program Overview

## Monday June 26, 2017

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:30 AM - 6:00 PM</td>
<td>Waldorf Room</td>
<td>Registration</td>
</tr>
<tr>
<td>7:00 AM – 8:45AM</td>
<td>Waldorf Room</td>
<td>Breakfast</td>
</tr>
<tr>
<td>8:15 AM - 8:30 AM</td>
<td>Williford Room</td>
<td><strong>Opening Ceremony</strong>: Lanju Zhang, Tony Cai</td>
</tr>
<tr>
<td>8:30AM - 8:45 AM</td>
<td>Williford Room</td>
<td><strong>Keynote I</strong>: Barry Nussbaum</td>
</tr>
<tr>
<td>8:45 AM - 9:45 AM</td>
<td>Williford Room</td>
<td><strong>Keynote II</strong>: Xiaoli Meng</td>
</tr>
<tr>
<td>9:45 AM - 10:00 AM</td>
<td></td>
<td>Break</td>
</tr>
<tr>
<td>10:00 AM - 11:40 PM</td>
<td>See program</td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>11:40 PM - 1:00 PM</td>
<td></td>
<td>Lunch on own</td>
</tr>
<tr>
<td>1:00 PM - 2:40 PM</td>
<td>See program</td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>2:40 PM - 2:50 PM</td>
<td></td>
<td>Break</td>
</tr>
<tr>
<td>2:50 PM - 4:30 PM</td>
<td>See program</td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>4:45PM - 5:45 PM</td>
<td>Williford Room</td>
<td><strong>30-year celebration session</strong></td>
</tr>
<tr>
<td>6:00 PM - 9:00 PM</td>
<td>Waldorf Room</td>
<td>Mixer</td>
</tr>
</tbody>
</table>

## Tuesday June 27, 2017

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:30 AM - 5:30 PM</td>
<td>Waldorf Room</td>
<td>Registration</td>
</tr>
<tr>
<td>7:00 AM - 8:45 AM</td>
<td>Waldorf Room</td>
<td>Breakfast</td>
</tr>
<tr>
<td>8:00 AM - 9:00 AM</td>
<td>Williford Room</td>
<td><strong>Keynote III</strong>: Rod Little</td>
</tr>
<tr>
<td>9:00 AM - 9:10 AM</td>
<td></td>
<td>Break</td>
</tr>
<tr>
<td>9:10 AM - 10:10 AM</td>
<td>Williford Room</td>
<td><strong>Keynote IV</strong>: Ram Tiwari</td>
</tr>
<tr>
<td>10:10 AM - 10:30 AM</td>
<td></td>
<td>Break</td>
</tr>
<tr>
<td>10:30 AM - 12:10 PM</td>
<td>See program</td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>12:10 PM - 1:30 PM</td>
<td></td>
<td>Lunch on own</td>
</tr>
<tr>
<td>1:30 PM - 3:10 PM</td>
<td>See program</td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>3:10 PM - 3:30 PM</td>
<td></td>
<td>Break</td>
</tr>
<tr>
<td>3:30 PM - 5:10 PM</td>
<td>See program</td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>6:30 PM - 9:30 PM</td>
<td></td>
<td><strong>Off site Banquet</strong> (30-year celebration performance)</td>
</tr>
</tbody>
</table>

## Wednesday June 28, 2017

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 AM - 1:00 PM</td>
<td>Waldorf Room</td>
<td>Registration</td>
</tr>
<tr>
<td>7:30 AM – 9:00 AM</td>
<td>Waldorf Room</td>
<td>Breakfast</td>
</tr>
<tr>
<td>8:30 AM - 10:10 AM</td>
<td></td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>10:10 AM - 10:30 AM</td>
<td></td>
<td>Break</td>
</tr>
<tr>
<td>10:30 AM - 12:10 AM</td>
<td></td>
<td><strong>Parallel Sessions</strong></td>
</tr>
</tbody>
</table>
Barry D. Nussbaum
President, American Statistical Association

Barry D. Nussbaum was the Chief Statistician for the U.S. Environmental Protection Agency from 2007 until his retirement in March, 2016. He started his EPA career in 1975 in mobile sources and was the branch chief for the team that phased lead out of gasoline. Dr. Nussbaum is the founder of the EPA Statistics Users Group. In recognition of his notable accomplishments he was awarded the Environmental Protection Agency's Distinguished Career Service Award. Dr. Nussbaum has a bachelor's degree from Rensselaer Polytechnic Institute, and both a master's and a doctorate from the George Washington University. In May, 2015, he was elected the 112th president of the American Statistical Association. He has been a fellow of the ASA since 2007. He has taught graduate statistics courses for George Washington University and Virginia Tech and has even survived two terms as the treasurer of the Ravensworth Elementary School PTA.

Title: Invitation to Improve the Statistical Profession: A Discussion of the Current ASA "Asian Initiative"
Location and Time: Williford Room, Monday June 26, 8:30AM - 8:45 AM
Xiao-Li Meng

Department of Statistics, Harvard University

Xiao-Li Meng, Dean of the Harvard University Graduate School of Arts and Sciences (GSAS), Whipple V. N. Jones Professor and former chair of Statistics at Harvard, is well known for his depth and breadth in research, his innovation and passion in pedagogy, and his vision and effectiveness in administration, as well as for his engaging and entertaining style as a speaker and writer. Meng has received numerous awards and honors for the more than 150 publications he has authored in at least a dozen theoretical and methodological areas, as well as in areas of pedagogy and professional development; he has delivered more than 400 research presentations and public speeches on these topics, and he is the author of "The XL-Files," a regularly appearing column in the IMS (Institute of Mathematical Statistics) Bulletin. His interests range from the theoretical foundations of statistical inferences (e.g., the interplay among Bayesian, frequentist, and fiducial perspectives; quantify ignorance via invariance principles; multi-phase and multi-resolution inferences) to statistical methods and computation (e.g., posterior predictive p-value; EM algorithm; Markov chain Monte Carlo; bridge and path sampling) to applications in natural, social, and medical sciences and engineering (e.g., complex statistical modeling in astronomy and astrophysics, assessing disparity in mental health services, and quantifying statistical information in genetic studies). Meng received his BS in mathematics from Fudan University in 1982 and his PhD in statistics from Harvard in 1990. He was on the faculty of the University of Chicago from 1991 to 2001 before returning to Harvard as Professor of Statistics, where he was appointed department chair in 2004 and the Whipple V. N. Jones Professor in 2007. He was appointed GSAS Dean on August 15, 2012.

Title: Personalized Treatment: Sounds heavenly, but where on Earth did they find the right guinea pig for me?

Location and Time: Williford Room, Monday June 26, 8:45 AM - 9:45 AM

Abstract: What data are relevant when making a treatment decision for me? What replications are relevant for quantifying the uncertainty of this personalized decision? What does "relevant" even mean here? The multi-
resolution (MR) perspective from the wavelets literature provides a convenient theoretical framework for contemplating such questions. Within the MR framework, signal and noise are two sides of the same coin: variation. They differ only in the resolution of that variation - a threshold, the primary resolution, divides them. We use observed variations at or below the primary resolution (signal) to estimate a model and those above the primary resolution (noise) to estimate our uncertainty. The higher the primary resolution, the more relevant our model is for predicting a personalized response. The search for the appropriate primary resolution is a quest for an age old bias-variance trade-off: estimating more precisely a less relevant treatment decision versus estimating less precisely a more relevant one. However, the MR setup crystallizes how the tradeoff depends on three objects: (i) the estimand which is independent of any statistical model, (ii) a model which links the estimand to the data, and (iii) the estimator of the model. This trivial, yet often overlooked distinction, between estimand, model, and estimator, supplies surprising new ways to improve mean squared error. The MR framework also permits a conceptual journey into the counterfactual world as the resolution level approaches infinite, where "me" becomes unique and hence can only be given a single treatment, necessitating the potential outcome setup. A real-life Simpson's paradox involving two kidney stone treatments will be used to illustrate these points and engage the audience.

This talk is based on the following three articles:

Roderick Little

Department of Biostatistics, University of Michigan, MI, USA

Roderick Little is a professor of biostatistics at the University of Michigan. Little received a PhD in statistics from Imperial College, London University in the United Kingdom. His current research interests involve analysis of data with missing values; analysis of repeated measures data with drop-outs; survey sampling, focused on model-based methods for complex survey designs that are robust to misspecification and compared to the resulting inferences to classical methods based on the randomization distribution; and applications of statistics to epidemiology, public health, psychiatry, sample surveys in demography and economics, and medicine. From 2010-12 he served as the inaugural Associate Director for Research and Methodology and Chief Scientist at the U.S. Census Bureau.

Title: Some recent developments in the analysis of data with missing values

Location and Time: Williford Room, Tuesday June 27, 8:00 AM - 9:00 AM

Abstract: Missing data are a common problem in public health research. Methods for handling this problem are briefly reviewed, including (a) pros and cons of different forms of likelihood inference, specifically maximum likelihood, Bayes and multiple imputation; (b) penalized spline of propensity models for robust estimation under the missing at random assumption, and comparisons with other doubly-robust approaches; and (c) subsample ignorable likelihood methods for regression with missing values of covariates. I'll also discuss two aspects of a recent National Research Council study on the treatment of missing data in clinical trials, namely how missing data impacts the choice of estimand, and sensitivity analysis for assessing departures from assumptions of the primary analysis.
Ram C. Tiwari, Ph.D.

Ram C. Tiwari, Ph.D. is the Director for Division of Biostatistics, CDRH, effective June 27, 2016. He joined FDA in April 2008 as Associate Director for Statistical Science and Policy in the Immediate Office, Office of Biostatistics, Office of Translational Sciences, CDER. Prior to joining FDA, he served as Program Director at National Cancer Institute, NIH, and as Professor and Chair, Department of Mathematics, University of North Carolina at Charlotte.

Dr. Tiwari received his MS and PhD degrees from Florida State University in Mathematical Statistics; he is a Fellow of the American Statistical Association, and an Elected Member of the International Statistical Institute. He is a past President of the International Indian Statistical Association. He has published 200+ research papers on a wide range of statistical topics. His current research interests include developing frequentist and Bayesian methods in clinical trials and pre-and-post market drug/device safety evaluation.

Title: Bayesian approaches for benefit-risk assessment with examples

Location and Time: Williford Room, Tuesday June 27, 9:10 AM - 10:10 AM

Abstract: An important aspect of the drug evaluation process is to have an integrated benefit-risk assessment to determine, using some quantitative measures, whether the benefit outweighs the risk for the target population. The subject-level benefit-risk response is a five-category random variable with cell counts following a multinomial distribution. Assuming that the cell probabilities follow a Dirichlet distribution, we develop a Bayesian approach for the longitudinal assessment of benefit-risk using the global measures proposed by Chuang-Stein et al. In a more generalized approach, a power prior is used through the likelihood function to discount the information from previous visits. For the subject-level benefit-risk assessment, the cell-probability of the subject, with respect to a reference category, is modeled, on the logarithmic scale, as a generalized linear model using a Dirichlet process as a prior. The model is applied to drug/device clinical trial datasets.

Keywords: Benefit-risk, Dirichlet Distribution, Dirichlet process, Power Prior, Model Selection
ASA Bio-pharmaceutical Awards

Xuan Zhou, University of North Carolina
— Title: Sequential Outcome-Weighted Multicategory Learning for Estimating Optimal Individualized Treatment Rules
— Time: Monday, June 26, 10:00 AM - 11:40 AM
— Session 32: Student session (I)

Linjun Zhang, University of Pennsylvania
— Title: CHIME: Clustering of High-Dimensional Gaussian Mixtures with EM Algorithm and Its Optimality
— Time: Wednesday, June 28, 8:30 AM - 10:10 AM
— Session 137: Statistical Inference for High-dimensional Linear Regression and Covariance Structure

Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper Award

Yi Deng, Emory University
— Title: Privacy-Preserving Methods for Horizontally Partitioned Incomplete Data
— Time: Monday, June 26, 2:50 PM - 4:30 PM
— Session 65: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences

ICSA Student Paper Awards

Haoran Li, UC Davis
— Title: An adaptable generalization of Hotelling's T2 statistic in high dimension
— Time: Tuesday, June 27, 1:30 PM - 3:10 PM
— Session 91: Student session (II)

Junwei Lu, Princeton
— Title: A Doubly Adaptive Inferential Method for Monotone Graph Invariants
— Time: Tuesday, June 27, 1:30 PM - 3:10 PM
— Session 91: Student session (II)

Cong Ma, Princeton
— Title: Inter-Subject Analysis: A Formal Theory
— Time: Tuesday, June 27, 1:30 PM - 3:10 PM
— Session 91: Student session (II)

Pulong Ma, U Cincinnati
— Title: Fused Gaussian Process for Very Large Spatial Data
— Time: Monday, June 26, 10:00 AM - 11:40 AM
— Session 32: Student session (I)

Xiaojun Mao, Iowa State
— Title: Matrix Completion with Covariate Information
— Time: Monday, June 26, 10:00 AM - 11:40 AM
— Session 32: Student session (I)
1. Multi-Regional Clinical Trials for Simultaneous Global New Drug Development

Instructors: Joshua Chen (Sanofi Pasteur), Hui Quan (Sanofi)

Course Length: half-day

Outline/Description:

Global clinical development strategy utilizing multi-regional clinical trials (MRCTs) plays a crucial role in developing innovative medicines. It is readily accepted that studying patients from many different regions within a single trial under a single protocol is an efficient method of trial design. The prevalence of these trials has been growing over the last few decades. MRCTs are most often conducted as a single trial focusing on the overall results, but when such trials are submitted to health authorities, the scope and concern often broaden to include the "local" results. Therefore, together with great opportunities, there are also tremendous challenges in the design, conduct and interpretation of MRCTs.

This short course is intended to provide a comprehensive view of the current status of conducting MRCTs in clinical development. We will start with the motivation, opportunities and challenges in simultaneous global clinical development using MRCTs. We then will focus on the design, monitoring and analysis/interpretation of MRCTs. These include the discussion of region definitions, the scenarios where different regions have differing requirements for a MRCT, the comparisons of models for MRCT, the quantifications of regional treatment effects, and methods for the assessment of consistency of regional treatment effects. Simulations and trial examples will be used to illustrate the ideas, methods and thinking. The registration fee for this short course includes a copy of the following book at a discounted rate.

Reference:


About the Instructors:

Dr. Yonghua (Joshua) Chen received his PhD in Statistics from the University of Wisconsin Madison. He is currently the Global Head of Biostatistics and Programming at Sanofi Pasteur. Before joining Pasteur, Josh worked on clinical development of small molecules, biologics and vaccines at Merck Research Laboratory Labs. His experience spans many therapeutic areas with major focus on human vaccines and antiviral drugs. He has extensive experience in study design, conduct and reporting of international clinical trials from proof-of-concept through regulatory approvals and life cycle management. His primary research interest is clinical trial designs, including group sequential methods, adaptive designs and multiregional clinical trials (MRCTs). He was a co-lead of the across-industry MRCT Consistency Working Group under PhRMA (2008-2011) and Society for Clinical Trials (2012-2014). He is a fellow of the American Statistical Association.

Dr. Hui Quan received his PhD degree in statistics from Columbia University in 1990. He is currently an associate VP and Global Head of the methodology group at the Biostatistics and Programming Department of Sanofi. He has 26 years of pharmaceutical industry experience in many therapeutic areas ranging from early phase to Phase IV studies. He has published 89 papers including 66 statistical papers. His research interests include multivariate analysis, safety analysis, multiplicity adjustment, missing handling, adaptive design, integrated data analysis, modeling and simulation, benefit/risk assessment and multi-regional clinical trials (MRCT). He served as a co-leader of the MRCT consistency assessment working group under PhRMA (2008-2011) and Society for Clinical Trials (2012-2014). He was elected as a fellow of the American Statistical Association in 2008.

2. Methods of biomarker and subgroup identification for personalized medicine

Instructor: Ilya Lipkovich (Advisory Services, QuintilesIMS)

Course Length: half-day

Outline/Description:

This short course will provide a description of a broad class of statistical methods dealing with exploratory subgroup analysis, i.e., subgroup search/biomarker discovery methods that can be applied both in early and late-phase clinical trials. We will also talk briefly about applying these methods in the context of non-randomized (observational studies).
Subgroup Analysis Working Group sponsored by the Society of Clinical Trials.

3. **Bayesian Adaptive Clinical Trials in the 21st Century**

**Instructor:** Ben Saville (Berry Consultants and Vanderbilt University Dept. of Biostatistics)

**Course Length:** half-day

**Outline/Description:**

As medical research continues to push into new frontiers of discovery and personalized patient care, it is imperative that clinical trial designs and statistical methodologies evolve to address the forthcoming challenges. In this course we discuss why Bayesian adaptive clinical trial designs are ideally suited to provide such innovation, with particular emphasis in phase II and III medical drug/device trials. Innovations include historical borrowing, hierarchical models, adaptive sample size with interim monitoring, Bayesian predictive probabilities, seamless transitions to subsequent phases, and multi-arm platform trials with a master protocol. I illustrate these methods using real clinical trial examples, and show the importance of simulations in navigating the regulatory environment. Hands-on small group activities are used to encourage discussion and learning.

**About the Instructor:**

Ben Saville is a Statistical Scientist for Berry Consultants, where he specializes in the design of innovative Bayesian adaptive clinical trials. He works primarily with medical device companies, pharmaceutical companies, and academic investigators to solve challenging problems via Bayesian designs, many of which are reviewed by the U.S. Food and Drug Administration (FDA). He is a frequent invited speaker at various statistical conferences, academic seminars, and lecture series, including short courses on adaptive clinical trial design. Dr. Saville earned his Ph.D. in Biostatistics from the University of North Carolina at Chapel Hill in 2008. Prior to joining Berry Consultants in 2014, he was an Assistant Professor of Biostatistics at Vanderbilt University School of Medicine where his methodological research focused on Bayesian hierarchical models, Bayesian adaptive clinical trials, and nonparametric methods for randomized clinical trials. At Vanderbilt, he collaborated extensively...
with medical researchers in the Department of Pediatrics and the Vanderbilt-Ingram Cancer Center, and was co-leader of an adaptive trials design workforce to promote innovative Bayesian methodology in clinical trials. In addition, he taught undergraduate and graduate courses in the Department of Biostatistics and Department of Biomedical Engineering. Dr. Saville has authored over 50 peerreviewed publications in the statistical and medical literature.

4. Statistical methods and software for multivariate meta-analysis

**Instructors:** Hai Tao Chu (Univ of Minnesota), Yong Chen (Upenn)

**Course Length:** half-day

**Outline/Description:**
This short course will provide an overview and tutorial of the cutting edge statistical methods and software for meta-analysis of diagnostic tests and multiple treatments comparisons. In particular, statistical models and methods will be presented in nontechnical format so that the course materials will be understood by applied statisticians. The short course will be presented with many case studies with detailed study background, well-formatted data sets, and relevant SAS and WinBUGS/R codes. Those details will enable audiences to understand the benefits of applying those innovative statistical methods developed recently, and to be able to apply those models in practice. Emphasis is given on the intuition behind those models and how to communicate the significance of applying those models with clients and customers. In return, those innovative methods and case studies will allow the audiences to achieve better estimation in meta-analysis of the accuracy of diagnostic tests and network clinical trials with multiple treatments.

**About the Instructors:**
Dr. Hai Tao Chu, Associate Professor of Biostatistics at the University of Minnesota. His methodology research focuses on comparative effectiveness research, meta-analysis and diagnostic test accuracy studies, and he has been working on meta-analysis for more than ten years. He has published over 140 peer-reviewed manuscripts, including over 30 on systematic reviews and meta-analysis in top ranked statistical and medical journals such as JASA, Biometrics, SIM, SMMR, Clinical Trials, JNCI, AIDS and AJE, and coauthored four R packages pcnetmeta, altmeta, xmeta and mmeta. Dr. Chu’s research on meta-analysis has been supported by FDA, AHRQ, NIAID, NIDCR and NLM. Dr. Chu is an ASA Fellow and elected member of the Society for Research Synthesis Methodology.

Dr. Yong Chen, Assistant Professor of Biostatistics at the University of Pennsylvania, has been working on the field of meta-analysis since 2008. He has published over 50 peer-reviewed manuscripts, including 15 on statistical methods for meta-analysis in top statistical journals such as Biometrics, JRSS-C, SIM and SMMR. He is also one of the main contributors of the R package for multivariate meta-analysis of binary data mmeta and package for multivariate meta-analysis that accounts for publication bias for continuous and binary data xmeta. His research has been supported by both AHRQ and NIH.

5. Clinical Trials for Time-to-Event Outcomes: Current Practice and New Developments

**Instructor:** Song Yang (NIH)

**Course Length:** half-day

**Outline/Description:**
Clinical trials assess the safety and effectiveness of new diagnosis, treatment, or prevention strategies. Very often the relevant outcome is a time-to-event outcome. While many time honored statistical tools, such as the log-rank test and Cox model inference, have been the cornerstones for time-to-event data analysis, many new methods have been developed in recent years. However, the systematic introduction of these new methods and their incorporation into clinical research practice has been rare. Adding to the complication and confusion is the fact that some approaches are more appropriate for exploratory studies while others are more appropriate for confirmative clinical research. The first part of the proposed short course would cover essential tools for clinical trials with time-to-event outcomes, such as sample size and power calculation, the log-rank test and its variations, interim analysis, and hazard ratio estimation. In the second part, we would go beyond the usual toolbox of the logrank test and Cox model inference and look at cutting-edge developments that improve and expand the
current practice, working with alternative models and measures for the outcomes to increase the flexibility, efficiency, applicability, and generalizability of clinical trial research. Depending on the audience, specific topics of this short course would include (i) clinical trial design, sample size and power calculation, (ii) the log-rank test and its variations, (iii) interim analysis, (iv) hazard ratio estimation, (v) time-dependent treatment effect, (vi) average hazard ratio and superiority probability, (vii) absolute risk reduction and restricted mean survival, (viii) covariate-adjusted analysis, and (ix) other measures and models. Applications to clinical trials, including some of the well-known large trials, will be illustrated using R so that the attendees will be able to understand the implementation procedures and analyze their own research data.

References:


About the Instructor:

Song Yang (Ph.D., Michigan State University, 1987) is a senior mathematical statistician and program officer at the Office of Biostatistics Research, National Heart, Lung, and Blood Institute, NIH. Before joining NIH, Dr. Yang was a full professor and the Statistics Coordinator at Texas Tech University. His major research interests are survival analysis and clinical trials. In the last 30 years, Dr. Yang has published numerous articles in methodology research and applied research in journals such as Ann. Statist., Biometrics, Biometrika, JASA, Stat. in Med., and NEJM. He is also a frequent reviewer for these journals and for National Science Foundation proposals, and has co-authored the R package YPmodel. Dr. Yang has given numerous invited presentations at major conferences and university seminars. He is an associate editor for Lifetime Data Analysis and Stat. & Prob. Lett., a guest lecturer of the Foundation for Advanced Education in the Sciences at NIH, and a past guest co-editor for Stat. in Med. Since joining NIH, Dr. Yang has served as the project office statistician of a dozen large-scale clinical trials and studies and has received the NIH Director’s Award and many other NIH awards for his service.

6. Integrative Multivariate Statistical Learning in Healthcare Research with Real-World Data

Instructor: Kun Chen (Univ of Connecticut), Dingfeng Jiang (AbbVie)

Course Length: half-day

Outline/Description:

This short course starts with an overview of recent problems arising from healthcare studies with large-scale heterogeneous data; examples include pragmatic trial, drug development, outcome research, suicide prevention, and opioid abuse. In these problems, a common scheme is that measurements of several distinct yet interrelated characteristics pertaining to a single set of subjects are collected from an array of disparate sources. For example, individual health data may come from insurance claims, pharmacy visits, clinical records, patient surveys, and government statistics. The availability of such complex data makes tackling many scientific and practical problems possible through “integrative statistical learning”, which is undergone exciting development and is pushing for a refinement of the conventional multivariate learning toolkit. In this short course, several classes of multivariate learning techniques for simultaneous dimension reduction, feature selection and model estimation will be introduced, together with discussions of several practical case studies in healthcare. The course consists of 4 modules: 1) overview of problems and statistical challenges in healthcare studies with big data; 2) integrative multivariate data reduction techniques with case studies; 3) integrative predictive modeling techniques with case studies; 4) recent developments on multi-view data fusion. The participants will have the opportunity to go through examples using newly developed R packages.

About the Instructors:
Dr. Kun Chen is an Assistant Professor in the Department of Statistics, University of Connecticut (UConn), and a Research Fellow at the Center for Public Health and Health Policy, UConn Health Center. Chen’s research interests include multivariate statistical learning, high-dimensional statistics, and health informatics with large-scale heterogeneous data. He has extensive interdisciplinary research experience in a variety of fields including insurance, ecology, biology, agriculture, medical imaging, and public health. Chen’s research projects have received funding from the National Institutes of Health, the Simons Foundation, the National Science Foundation, etc. Currently he is involved in a data-driven suicide prevention study through integrating big data from disparate sources. Chen was a Co-Editor of the 2015 ICSA Symposium Proceeding Book, and serves as an Associate Editor of Sankhya: The Indian Journal of Statistics since 2016. He was recognized for Teaching Excellence at UConn for multiple times.

Dr. Dingfeng Jiang is a statistical manager at AbbVie Inc. Jiang’s research interest include high-dimensional statistics, variable selection, and causal inference in observational studies. He has extensive research experience in designing observational outcome research using big healthcare data, with application in diabetes, oncology and immunology therapy areas. He has served as reviewers for multiple statistical journals. He is an editorial board member for Heliyon, an open access journal published by Elsevier.

7. Phase II Clinical Development

Instructors: Naitee Ting (Boehringer Ingelheim), Shuyen Ho (PAREXEL)

Course Length: half-day

Outline/Description:

Clinical development of new drugs or new biologics can be broadly divided into four phases - Phases I, II, III, and IV. Phase I is primarily to study the pharmacokinetics (PK) properties and to estimate maximally tolerable dose (MTD) and Phase II is for proof of concept (PoC) and dose ranging. Phase III is designed and executed for registration and Phase IV is for post marketing purposes. Because product efficacy is not the focus of Phase I, Phase II is typically the first time drug efficacy is tested in patients with the targeted disease.

After Phase II, a major decision will be made for the progressing of test product into Phase III or not. Therefore Phase II can be considered as the most critical phase in clinical development. This course emphasizes the importance and challenges of Phase II development, examples and case studies are illustrated to provide guidance on Phase II study designs and analysis methods.

About the Instructors:

Naitee Ting is a Fellow of ASA. He is currently a Director in the Department of Biostatistics and Data Sciences at Boehringer-Ingelheim Pharmaceuticals Inc. (BI). He joined BI in September of 2009, and before joining BI, he was at Pfizer Inc. for 22 years (1987-2009). Naitee received his Ph.D. in 1987 from Colorado State University (major in Statistics). He has an M.S. degree from Mississippi State University (1979, Statistics) and a B.S. degree from College of Chinese Culture (1976, Forestry) at Taipei, Taiwan. Naitee published articles in Technometrics, Statistics in Medicine, Drug Information Journal, Journal of Statistical Planning and Inference, Journal of Biopharmaceutical Statistics, Biometrical Journal, Statistics and Probability Letters, and Journal of Statistical Computation and Simulation. His book “Dose Finding in Drug Development” was published in 2006 by Springer, and is considered as the leading reference in the field of dose response clinical trials. The book “Fundamental Concepts for New Clinical Trialists”, coauthored with Scott Evans, was published by CRC in 2015. Naitee is an adjunct professor of Columbia University, University of Connecticut and University of Rhode Island. Naitee has been an active member of both the American Statistical Association (ASA) and the International Chinese Statistical Association (ICSA).

Shuyen Ho is a Biostatistics Director at PAREXEL International in Durham, North Carolina. Shuyen has worked in the pharmaceutical industry for 27 years. Prior to PAREXEL, he was a Clinical Statistics Director at GlaxoSmithKline (GSK) and Group Leader at Merck. He specializes in Phase II & III clinical development and has helped developed widely used respiratory medicines such as Claritin, Advair and Veramyst. Shuyen received his PhD in Statistics from University of Wisconsin – Madison, and his Bachelor in Applied Mathematics from Taiwan.
8. Principles of Multiple Comparisons, with Applications

Instructors: Jason C. Hsu (OSU), Frank Bretz (Novartis), Yi Liu (Tekeda), Dong Xi (Novartis)

Course Length: half-day

Outline/Description:

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 prespecified 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.

References:


About the Instructors:

Jason Hsu is an emeritus professor in the Department of Statistics of the Ohio State University. His recent research on multiple comparisons focuses on efficacy inference in subgroups and their mixtures, both in discovery studies and in confirmatory studies.

Frank Bretz joined Novartis in 2004, where he is currently Global Head of the Statistical Methodology and Consulting group. He has supported the methodological development in various areas of drug development, including dose-finding, multiple comparisons, and adaptive designs.

Yi Liu earned her PhD from department of statistics at the Ohio State University in 2009. She currently works as a senior manager in the department of global statistics and programming at Takeda Pharmaceuticals Inc. Her research interests include adaptive designs and multiple comparisons.

Dong Xi joined Novartis in 2013, where he is an Expert Statistical Methodologist in the Statistical Methodology and Consulting group. His research interests include multiple
comparisons and dose-finding. Prior to his current role, he got his PhD in statistics from Northwestern University.

9. Bayesian Adaptive Phase I Oncology Trials: Methodology and Implementation methodology

Instructors: Satrajit Roychoudhury (Novartis), Beat Neuenschwander (Novartis)

Course Length: half-day

Outline/Description:

Phase I trials in Oncology are usually small adaptive dose-escalation trials. The aim is to approximately understand the dose-toxicity profile of a drug, and, eventually, to find a reasonably safe dose for future testing. A lot of statistical research for Phase I trials has accumulated over the past 25 years, with modest impact on statistical practice. The vast majority of trials still follow the 3+3 design, despite the fact that it often misses the targeted dose (poor operating characteristics) and fails to provide a real understanding about true toxicity rates (no statistical inference).

In this course we present a comprehensive and principled statistical approach. The implementation is Bayesian, with the following main parts: a parsimonious model for the dose–toxicity relationship; the possibility to incorporate contextual information (“historical data”) via priors; and, safety-centric metrics (overdose probabilities) which inform dose adaptations under appropriate overdose control.

After some basic clinical and statistical considerations, we introduce the statistical methodology for the single-agent setting, and then extend it to dual- and triple-combinations. Applications and a discussion about implementation (such as basic WinBUGS code) issues complement this training and provide practical insights into Phase I trials.

References:


About the Instructors:

Dr. Satrajit Roychoudhury is Director and member of Statistical Methodology and Consulting at Novartis Pharmaceuticals, USA. Before joining Novartis in 2009, he was at Schering Plough Research Institute. He has a strong expertise in interacting with statisticians, clinicians and scientists regarding implementation of innovative statistical methodology in clinical trial. He has extensive experience of development and implementation of model based approaches in Phase I Oncology trials. He co-authored several publications/book chapters in this area. He also provided half-day training on similar topic at Joint Statistical Meeting, Seattle and ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop, Washington D.C. in 2015. In addition he has 5 years of extensive experience in working with different phase clinical trials. His area of research includes the use of Bayesian methods in clinical trials, especially evidence synthesis. He authored several publications in peer reviewed journals and served as a referee for many journals. He is currently working on and is currently working on the book “Statistical Challenges in Oncology Drug Development” (CRC Press).

Dr. Beat Neuenschwander received his Ph.D. in 1991 from the University of Bern, Switzerland. From 1992 to 2000, he had a joint appointment as a consultant and statistical analyst at the Department of Social and Preventive Medicine, University of Berne, and the Department of Epidemiology, Swiss Federal Office of Public Health. He joined Novartis Pharmaceuticals in 2001, where he worked as a trial statistician, statistical modeler, and methodologist. He joined the Novartis Oncology Business Unit as a Biometrical Fellow in 2010. Beat Neuenschwander has contributed significantly to the methodological development in various
areas of drug development, including the design of Bayesian phase I cancer trials, and the use of historical data in Phase I and II clinical trials. During the past 20 years in academia and industry, he has given various Bayesian trainings. He served as an associate editor for Biometrical Journal. He has authored or co-authored 30 articles in peer-reviewed journals, and is currently working on the book “Bayesian Reasoning in Drug Development” (Springer) with his colleague Heinz Schmidli (Novartis Pharma).

10. Design and Analysis of Real-world Clinical Studies Using SAS and R

Instructor: Macaulay Okwuokenye (Biogen)

Course Length: half-day

Outline/Description:
Although randomized, controlled, double blind experiments (RCDBE) are considered the gold standard for evidentiary inference, many instances exist where a RCDBE may be unethical or impractical. Moreover, a RCDBE may not be reflective of real-world settings. Such instances may warrant reliance on data from studies from more practical designs to support decision and policy making—e.g., strengthening comparative effectiveness profile of a therapeutic product or medical device using data from routine clinical practice or disease registries. The design, conduct, and statistical analyses of departures from the RCDBE can and should mirror randomized counterpart. Departures could include non-randomized, non-controlled, non-double blinded, or any combination of these. Participants will learn methods for designing credible real-world clinical studies that mimic traditional randomized trials, and statistical analysis of data therefrom. These methods derive from causal inference framework, including propensity score (matching, inverse probability of treatment weight, stratification by propensity score), genetic matching, etc. Best methodological practices for improving credibility of findings will be discussed. Software implementation of these methods using SAS and R will be demonstrated.

About the Instructor:
Macaulay Okwuokenye is a Principal biostatistician at Biogen Inc., MA and an adjunct faculty of biostatistics at Jiane-Ping Hsu College of Public Health (JPHCOPH), Georgia Southern University, GA. At Biogen, he supports design, conduct, and analyses of clinical trials and real-world clinical studies and exploratory data analysis. He has years of academic research and several years of clinical research, statistical consultation, and pharmaceutical and biotechnology industry experience spanning phase II, IV, registries, real-world, and bioequivalence studies. Macaulay is adept in evidentiary inference for real-world data and comparative effectiveness of therapeutic products. He received the masters and doctoral degree in biostatistics from the JPHCOPH, Georgia Southern University. He has published in methodological and applied international journals and is a reviewer for international statistical journals. He is a speaker at international conferences, statistical workshop, and academic seminars.

11. Experiences and Case Studies in Adaptive Clinical Trial Design

Instructors: Shiowjen Lee (FDA), Min (Annie) Lin (FDA)

Course Length: Half-day

Outline/Description:
For the past decades, there has been considerable interest among pharmaceutical and other medical product developers in adaptive design clinical trials, in which knowledge learned from data of an ongoing trial affects design features or analysis of the study. Following the release of the FDA draft Guidance document on adaptive design clinical trials in early 2010, there have been high expectations of an increase in regulatory submissions involving adaptive design features, particularly for confirmatory trials. Additionally, the recently passed 21st Century Cures Act encourages a broader use of adaptive design in clinical studies. Despite all this, there remain some concerns and questions regarding statistical analyses and operational challenges in conducting adaptive design clinical trials. We will share our experiences in the review of adaptive design proposals, including surveys performed regarding regulatory submissions of adaptive design proposals as well as case studies which have been reviewed. We will also provide general recommendations for developing proposals for such trials. Our motivation in instructing this short course is to encourage the best study design proposals to be submitted to FDA. Sometimes these designs can be adaptive and sometimes a simpler design is most efficient.
About the Instructors:

Shiowjen Lee, Ph.D. is currently Team Leader in Center for Biologics Evaluation and Research (CBER), US FDA. In her current role, Dr. Lee oversees regulatory submissions supporting the biologic products of tissue, cellular and gene therapies regulated by Office of Tissue and Advanced Therapies (OTAT). Prior to the current role, she has worked at Center for Drug Evaluation and Research (CDER) and Center for Devices and Radiological Health (CDRH) as a Mathematical Statistician in reviews of regulatory submissions. In her work, she has been reviewing a wide range of regulatory applications including cancer vaccines, anti-inflammatory products, products (including biologic and devices) to treat cardiovascular disease, rare disease and many others. She has expertise in both statistics and clinical trial design and has broad experiences in products regulated at different Centers in FDA. Dr. Lee’s work in FDA has also included developing regulatory guidance and policies regarding specific indications, and statistical issues. She has a great ability collaborating and communicating statistics to review teams and non-statisticians. Dr. Lee has been a recipient for numerous awards at Center and FDA levels in FDA including Outstanding Service award, Team Excellence award, Center Director’s Award for Improving Communication, Managerial Excellence award, and many others. She was a Co-Chair of 2014 FDA/Industry Statistics Workshop that was held at Washington DC.

Min (Annie) Lin, Ph.D. joined the Division of Biostatistics at Center for Biologics Evaluation and Research (CBER), US FDA in 2011 as a Mathematical Statistician with special focus on Adaptive Designs. In her current role, Dr. Lin performs critical statistical reviews of regulatory premarketing submissions to the Office of Tissues and Advanced Therapies (OTAT) and participates in regulatory guidance development. She supervises the FDA interns in conducting biostatistical methodology researches in the area of adaptive designs that are related to CBER regulated biologics products and medical devices. Prior to joining the FDA, Dr. Lin was an Assistant Professor in the Department of Biostatistics and Bioinformatics at Duke University School of Medicine. She served as statistical investigator/co-investigator in various pharmaceutical funded and government-funded studies for a wide range of therapeutic areas.

12. Quantitative Sciences for Safety Monitoring during Clinical Development


Course Length: Full-day

Outline/Description:

In an effort to better promote public health and protect patient safety, there is growing interest in developing a systematic approach for safety evaluation of pharmaceutical products, not only for post-marketing safety surveillance, but also for pre-marketing safety monitoring. Recent regulatory guidance, such as CIOMS VI, ICH E2C and FDA IND safety reporting guidance (2012, 2015), have highlighted the importance and given recommendations on aggregate safety monitoring. Biostatisticians and other quantitative scientists can closely engage with clinical and regulatory scientists and play a vital role in these efforts. In 2015, to better enable this, the ASA Biopharmaceutical Section established a working group on clinical safety monitoring.

This tutorial session will present the work that has been done by this ASA safety monitoring working group, in two parts: 1. Summary of relevant regulatory guidance and results of an industry-wide survey on current process and technology enablement. 2. Discussion of various statistical methods for safety monitoring, these include blinded vs unblinded analyses, Frequentist vs Bayesian approaches, premarketing vs post marketing strategies, static vs dynamic assessments, trial-level vs program-level data aggregations, as well as visual analytical methods for safety data monitoring.

A continual theme throughout the day will be the opportunity for quantitative scientists to engage in understanding the safety challenges and help provide solutions for colleagues in the broader safety surveillance fields. Audience participation will be highly encouraged.

About the Instructors:

Susan Duke joined AbbVie in early 2016 as director and head, Safety Statistics in Data Science and Statistics. Beginning at Genentech in 1986, she held positions at small
biotech/device companies, followed by 18 years at GlaxoSmithKline, completing her tenure there in BenefitRisk Evaluation in the Drug Safety department. In her 30th year in industry, her most notable achievements have been in safety graphics, quantification of benefit-risk processes and deliverables, and a track record for successfully implementing new ways of working across multiple disciplines.

Krishan Singh has been working in the pharmaceutical industry for 29 years, starting at Smith Kline & French which subsequently became GlaxoSmithKline following M&A. In his 29 years in the industry as a statistician, he has supported clinical development of drugs across a number of therapy areas leading to regulatory submissions and market authorization of six new drugs in cardiovascular, inflammation and tissue repair, anti-infectives, respiratory and rare diseases.

Krishan brings extensive experience in the application of statistical methodologies for the evaluation of safety and efficacy of investigational drugs.

Wenquan Wang has been working in the clinical development for 13 years, mainly in oncology. He has worked for GlaxoSmithKline, Eisai, and currently Morphotek. He has worked on chemotherapeutic agents, cancer supportive care, and monoclonal antibodies as add-on to chemo-backbone or maintenance therapy, in which benefit risk balance is always a critical point in evaluating these drugs. Wenquan has particular interest in evaluating anti-drug-antibodies when patients are treated with monoclonal antibodies. Ed Whalen has worked in the biopharma industry for 25 years, six at Bayer Inc and 19 at Pfizer Inc. He has worked in anti-infectives, oncology, cardiovascular, neuroscience, and pain. For the last 15 years his work has focused on neuroscience and pain. Because of the variability in patient efficacy response, Ed has ventured into safety statistics to better understand the benefit risk balance for these drugs and in general.
# ICSA 30-Year Celebration Panel Session

**ICSAR 30**  
**Year Celebration Panel Session**

---

**1987-2017**

**Session Organizer:** Mei-Ling Ting Lee, U Maryland  
**Session Chair:** Tony Cai, U Pennsylvania

**Panelists:**

<table>
<thead>
<tr>
<th>Service Year</th>
<th>Name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987-1988</td>
<td>George C. Tiao</td>
<td>U Chicago</td>
</tr>
<tr>
<td>1989</td>
<td>James C. Fu</td>
<td>U Manitoba</td>
</tr>
<tr>
<td>1990</td>
<td>Jia-Yeong Tsay</td>
<td>Retired Pharmaceutical Executive</td>
</tr>
<tr>
<td>1991</td>
<td>Grace Yang</td>
<td>U Maryland</td>
</tr>
<tr>
<td>2001</td>
<td>Chao Agnes Hsiung</td>
<td>National Health Research Institute</td>
</tr>
<tr>
<td>2002</td>
<td>William W. S. Wei</td>
<td>Temple U</td>
</tr>
<tr>
<td>2004</td>
<td>Frank Shen</td>
<td>Abbvie</td>
</tr>
<tr>
<td>2005</td>
<td>Jiahua Chen</td>
<td>U British Columbia</td>
</tr>
<tr>
<td>2008</td>
<td>Jane-Ling Wang</td>
<td>UC Davis</td>
</tr>
<tr>
<td>2010</td>
<td>Xuming He</td>
<td>U Michigan</td>
</tr>
<tr>
<td>2011</td>
<td>Naisyin Wang</td>
<td>U Michigan</td>
</tr>
<tr>
<td>2012</td>
<td>Ivan S. F. Chan</td>
<td>Abbvie</td>
</tr>
<tr>
<td>2013</td>
<td>Ming-Hui Chen</td>
<td>U Connecticut</td>
</tr>
<tr>
<td>2014</td>
<td>Ying Lu</td>
<td>Stanford U</td>
</tr>
<tr>
<td>2015</td>
<td>Wei Shen</td>
<td>Lilly</td>
</tr>
<tr>
<td>2018</td>
<td>Aiyi Liu</td>
<td>NIH</td>
</tr>
</tbody>
</table>

**Time:** 4:45-5:45 pm, Monday, June 26, 2017  
**Venue:** Williford Rooms
Banquet Program

Menu

西湖牛肉羹  West Lake Beef Broth
椒盐虾  Pepper-and-Salt Fried Shrimp
荷芹酱爆三鲜  Sauteed Mixed Seafood
京都排骨  Braised Sparerib
金银蛋扒豆苗  Garden Pea Sprout with Egg
片皮鸭  Peking Duck
蒜蓉蒸龙虾  Steamed Lobster with Chopped Garlic
皇子菇牛肉  Beef with King Oyster Mushroom
清蒸盲曹  Steamed Seabass
冬菇伊面  Soft Noodles with Dried Mushroom
甜品  Dessert

Performance

Dance: Water Sleeves Dance (Xin Wang)
Miss You Suddenly (Li Wang, keyboard: Su Chen)
Thousand Miles Away  (Zhenzhen Zhang & Li Wang, keyboard: Su Chen)
Dance: The Blossom (Dora Guo)
Mi Manchelai (Zhenzhen Zhang, keyboard: Su Chen)

6:30-9:30pm June 27, 2017
Furama Restaurant  富丽华海鲜大酒楼
2828 S. Wentworth Ave, Chicago, IL 60616
Tel: 312-225-6888
Announcement: 2018 ICSA Applied Statistics Symposium

The 2018 ICSA Applied Statistics Symposium will be held from Thursday June 14 to Sunday June 17, 2018 at Hyatt Regency New Brunswick, New Brunswick New Jersey. You are cordially invited to the symposium!

Call for Invited Session Proposals

We welcome your session proposals. The invited session proposals will be reviewed and selected by the program committee. An invited session will be 100 minutes with four speakers or three speakers and one discussant. A proposal should include: 1) Session title; 2) Organizer(s) (with affiliation and email contact information); 3) Session chair (with affiliation and email contact information); 4) a list of speakers or discussant (with affiliation and email contact information and tentative talk titles). There is one-talk policy, but one can serve as a discussant in another invited session while speaking in an invited or contributed session. It is required to confirm all speakers and discussants availability before submitting a proposal. The submission deadline is December 1, 2017.

Call for Course Proposals

There will be six short course slots in the symposium. We welcome your submission of short course proposals. Please include in your proposal a brief description of the course with title, course plan, target audience and estimated audience size, plus a short biography of the instructor(s). The short course proposals will be reviewed and selected by the program committee. The submission deadline is January 1, 2018.

Call for Student Paper Award Applications

We welcome students to submit their original research article for the competition for a Student Paper Award. Up to eight student award winners will be selected. Each winner will receive a plaque or certificate, an award for travel and registration reimbursement up to $1,000 or a cash award of $550, whichever is bigger, as well as a free registration for a short course. The deadline for application is March 1, 2018.

Call for Sponsorship:

Corporate support has always been and will continue to be a critical factor to the success of the Symposium and ICSA. The 2018 ICSA Applied Statistics Symposium is seeking for sponsorship to support student awards, reimburse the keynote speakers and other symposium costs. Sponsors at both Gold and General levels will be recognized through our publications and onsite signage and reach a broad domestic and international audience through the symposium. At the Gold level, sponsors will also be given the opportunities to set up an exhibition booth and a placement service at the symposium. Your corporate sponsorship will not only demonstrate your leadership in the industry but also the corporation's recognition of statisticians' significant contribution in the business. Any level of contributions will be appreciated.

2018 ICSA Applied Statistics Symposium Organizing Committee
JOIN US!

2017 ICSA Midwest & NIC-ASA Joint Fall Meeting
A Great Event for Networking and Updating Quantitative Skills

When: October, 2017
Where: Northern Chicago, IL
Who: Statisticians, Business Analysts, Data Scientists and All Quantitative Scientists!

- Two well-known industry keynote speakers
- Day 1 short courses and day 2 parallel sessions
- Breakfast and lunch are included
- Generous student poster awards
- Career services for students available (including Citadel!)
- Free for student attendees who submitted posters
- New this year: Cocktail Reception for Everyone!!
Sunday, June 25. 1:00 PM - 2:40 PM

Session 1: Recent progress in high dimensional data analysis (Invited)
Room: CONFERENCE ROOM 5G, FIFTH FLOOR
Organizer: Cao, Hongyuan; University of Missouri.
Chair: Wenxin Zhou; Princeton University.
1:00 PM FDR control procedure integrating prior correlation structure with application to microbiome data
• Jun Chen1, Hongyuan Cao2 and Jian Xiao1. 1Mayo Clinic
2University of Missouri
1:25 PM Robust Factor Models with Covariates
Jiaying Fan1, Yuan Ke2 and Yuan Liao3. 1Princeton University
2Rutgers University
1:50 PM Variable and Group Selection with Prior Information
• Yuan Jiang1, Li Kai1, He Yuxiao2 and Zhang Heping3. 1Oregon State University
2Nielsen Company
3Yale University
2:15 PM Testing for Trends in High-dimensional Time Series
• Likai Chen and Wei Biao Wu. University of Chicago
2:40 PM Floor Discussion.

Session 2: Spatial and Temporal Modeling for Environmental Data (Invited)
Room: CONFERENCE ROOM 5E, FIFTH FLOOR
Organizer: Chang, Won; University of Cincinnati.
Chair: Chang, Won; University of California.
1:00 PM Interannual variation in the characteristics of precipitation events under a changing climate
• Chen Chen1, Won Chang2, Wenwen Kong3, Jiali Wang4, V. Kao Kotamarthi5, Michael L. Stein1 and Elisabeth J. Moyer1. 1University of Chicago
2University of Cincinnati
3University of California at Berkeley
4Oregon State University
5Newcastle University
1:25 PM Statistics-Based Compression of Global Wind Fields
• Jaehong Jeong1, Stefano C广汽 cocoa2, Paola Crippa2 and Marc Genton1. 1King Abdullah University of Science and Technology
2Newcastle University
1:50 PM Approaches for Massively Spatial Data and Applications in Remote Sensing
• Emily Kang and Pulong Ma. University of Cincinnati
2:15 PM Modeling Precipitation Extremes using Log-Histospline
• Whitney Huang1, Doug Nychka2 and Hao Zhang1. 1Purdue University
2National Center for Atmospheric Research
2:40 PM Floor Discussion.

Session 3: Recent advances in statistical methods for brain connectome networks (Invited)
Room: CONFERENCE ROOM 5H, FIFTH FLOOR
Organizer: Chen, Shuo; University of Maryland.
Chair: Chen, Shuo; University of Maryland.
1:00 PM Estimating Brain Pathway Effects Using Large-scale Multilevel Models
• Xi Luo and Yi Zhao. Brown University
1:25 PM Robust Graph Change-point Detection for Brain Evolution Study
• Fang Han1, Xi Chen2, Honglang Wang3, Lexin Li4 and Brian Caffo5. 1University of Washington
2New York University
3Indiana University-Purdue University Indianapolis
4University of California, Berkeley
5Johns Hopkins University
1:50 PM A Network-Object Method to Uncover Hidden Disease-Related Brain Connectome
Shuo Chen. University of Maryland, College Park
2:15 PM Floor Discussion.

Session 4: New Bayesian methods in applied statistics (Invited)
Room: CONFERENCE ROOM 5F, FIFTH FLOOR
Organizer: Chen, Yi-Hau; Academia Sinica Taiwan.
Chair: Yun Chan Chi; National Cheng Kung University.
1:00 PM Detecting copy number variations from next-generation sequencing data via a Bayesian procedure
• Guan-Hua Huang1 and Yu-Chung Wei2. 1National Chiao Tung University
2Feng Chia University
1:25 PM Estimating links of a network from time to event data
• Tso-Jung Yen1, Zong-Rong Lee2, Yi-Hau Chen3, Yu-Min Yen4 and Jing-Shiang Hwang5. 1Institute of Statistical Science, Academia Sinica
2Institute of Sociology, Academia Sinica
3National Chengchi University
1:50 PM Bayesian Empirical Likelihood Based Analysis for Patterned Missing data.
2:15 PM Floor Discussion.

Session 5: Advance in Statistical Methods for Large and Complex Data (Invited)
Room: CONFERENCE ROOM 5I, FIFTH FLOOR
Organizer: Kong, Dehan; University of Toronto.
Chair: Kong, Dehan; University of Toronto.
1:00 PM Aggregating Information from Unlabeled Heterogenous Sources
Haoyang Liu and Chao Gao. University of Chicago
1:25 PM High Dimensional Propensity Score via Covariate Balancing
• Yang Ning1, Sida Peng2 and Kosuke Imai2. 1Cornell University
2Princeton University
Session 7: New methods development in Genetic and genomic data analysis (Invited)
Room: CONFERENCE ROOM 5B, FIFTH FLOOR
Organizer: Liu, Jin; Duke-NUS Singapore.
Chair: Yang, Can; Hong Kong Baptist University.
1:00 PM IMAC: A Statistical Framework for Integrating Multiple Annotations to Characterize Functional Roles
Can Yang. Hong Kong Baptist University
1:25 PM Robust Genetic Prediction of Complex Traits with the Latent Dirichlet Process Regression Models
Ping Zeng and Xiang Zhou. University of Michigan
1:50 PM SynthEx: A synthetic-normal based DNA sequencing tool for copy number alteration detection
Mengjie Chen1, Grace O. Silva2, Marni B. Siegel1, Lisle E. Mose2, Joel S. Parker2, Wei Sun3 and Charles M. Perou2, 1The University of Chicago, 2University of North Carolina - Chapel Hill, 3Fred Hutchinson Cancer Research Center
2:15 PM SSREM: A Summary-Statistics-based Random Effect Model to Estimating Heritability, Co-heritability
Jin Liu. DUKE-NUS Medical School
2:40 PM Floor Discussion.
Sunday, June 25. 3:15 PM - 4:55 PM

Session 10: Recent advances in methods for EHR based research (Invited)
Room: CONFERENCE ROOM 5G, FIFTH FLOOR
Organizer: Chen, Yong; University of Pennsylvania.
Chair: Duan, Rui; University of Pennsylvania.

3:15 PM Constructing stabilized dynamic treatment regimes using electronic health record data
Yingqi Zhao. Fred Hutchinson Cancer Research Center

3:40 PM Constructing Dynamic Treatment Regimes in Infinite-Horizon Settings
Ashkan Ertefaie. University of Rochester

4:05 PM Predicting complex phenotypes in electronic health records with semiparametric CCA

Dennis Agnieszka and Tianxi Cai. 1RAND Corporation
2Harvard University

4:30 PM Bias reduction methods for EHR data based association studies
Yong Chen. University of Pennsylvania

4:55 PM Floor Discussion.

Session 11: Recent Advances in High-dimensional Inference (Invited)
Room: CONFERENCE ROOM 5E, FIFTH FLOOR
Organizer: Wang, Lan; University of Minnesota.
Chair: Lan Wang; University of Minnesota.

3:15 PM Regression coefficients clustering in high dimensional data
Junghi Kim1, Hongtu Zhu1 and Xiao Wang2. 1University of Texas MD Anderson Cancer Center 2Purdue University

3:40 PM High-dimensional inference robust to sparsity
Jelena Bradic and Yinchu Zhu. University of California, San Diego

4:05 PM Homogeneity Test of Covariance Matrices with High-Dimensional Longitudinal Data
Ping-Shou Zhong1 and Runze Li2. 1Michigan State University 2Penn State University

4:30 PM A New Scope of Penalized Empirical Likelihood with High-Dimensional Estimating Equations
Jinyuan Chang1, Cheng Yong Tang2 and Tong Tong Wu2. 1Southwestern University of Finance and Economics 2Temple University 3University of Rochester

4:55 PM Floor Discussion.

Session 12: Survival Analysis and Genetics (Invited)
Room: CONFERENCE ROOM 5F, FIFTH FLOOR
Organizer: Fan, Razong; Georgetown University.
Chair: Chen, Wei; Children’s Hospital of Pittsburgh.

3:15 PM Selection models for enhancing power of tests of genetic associations in family studies
Richard Cook. University of Waterloo

3:40 PM Assessment of familial genetic cancer risk in the presence of competing risks
Lakhal-Chaieb M Hamed Lajmi. Laval University

4:05 PM An additive hazards model for gene level association analysis of survival traits of complex disorder
Chi-yang Chiu1, Richard Cook2 and Ruzong Fan3. 1National Institutes of Health. 2University of Waterloo, 3Georgetown University Medical Center

4:30 PM Pedigree-based Association Analysis of Survival Traits via Functional Regressions
Ruzong Fan. Georgetown University Medical Center

4:55 PM Floor Discussion.

Session 13: Design and Analysis of Dose Finding Clinical Trials for Combination Therapy (Invited)
Room: CONFERENCE ROOM 5I, FIFTH FLOOR
Organizer: Gao, Jingjing; AbbVie Inc.
Chair: Zhang, Hongtao; AbbVie Inc.

3:15 PM Evaluation of false positive rate based on E-R analyses for two compounds in fixed-dose product
Yaning Wang and Hao Zhu. U.S. FDA

3:40 PM Dose-Finding for Immunotherapy Combinations using a Conditionally Autoregressive Model
Thomas Braun. University of Michigan Department of Biostatistics

4:05 PM Designing early phase drug combination trials in practice
Ying Yuan. MD Anderson Cancer Center

4:30 PM Modelling semi-attributable toxicity in dual-agent phase I trials
Graham Wheeler, Michael Sweeting, Adrian Mander, Shing Lee and Ken Cheung. 1Columbia University

4:55 PM Floor Discussion.

Session 14: Understanding the Microbiome Complexity –Genetics and Networks (Invited)
Room: CONFERENCE ROOM 5A, FIFTH FLOOR
Organizer: Li, Hongzhe; University of Pennsylvania.
Chair: Jing Ma; University of Pennsylvania.

3:15 PM Microbial network estimation using bias-corrected graphical lasso
Duo Jiang, Yuan Jiang and Thomas Sharpton. Oregon State University

3:40 PM Compositional mediation model for microbiome study
Michael Sohn and Hongzhe Li. University of Pennsylvania

4:05 PM A Testing Framework for Detecting Differential Microbial Networks
Jing Ma1, Yin Xia2, Hongzhe Li1 and T. Tony Cai3. 1University of Pennsylvania 2Fudan University

4:30 PM Estimate Overall Contribution and Build Risk Prediction Models in Large Scale Prospective Microbiome Studies
Jianxi Shi. National Cancer Institute

4:55 PM Floor Discussion.
Session 15: Recent Developments in Neuroimaging Analysis
(Invited)
Room: CONFERENCE ROOM 5H, FIFTH FLOOR
Organizer: Li, Lexin; UC Berkeley.
Chair: Miheye Ahn; University of Nevada, Reno.

3:15 PM  FMRI Preprocessing Changes the Statistical Properties of Your Data
*Daniel Rowe*, Linglong Kong* and Haochang Shou*. Marquette University
University of Alberta University of Pennsylvania

3:40 PM Tensor approximation of partial quantile regression for neuroimaging data analysis
*Dengdeng Yu*, Linglong Kong and Ivan Mizera. University of Alberta

4:05 PM Modeling dynamic brain response to pain progression using multilevel functional regression

4:30 PM Floor Discussion.

Session 16: Prediction and inference for high-dimensional data and nonparametric models (Invited)
Room: CONFERENCE ROOM 5B, FIFTH FLOOR
Organizer: Ma, Shujie; University of California Riverside.
Chair: Ma, Shujie; University of California Riverside.

3:15 PM Generalized Fiducial Inference for High-Dimensional Sparse Additive Models
Qi Gao*, Randy Lai* and Thomas Lee*. UC Davis U of Maine

3:40 PM Exploration of Large Networks via Fast and Universal Latent Space Model Fitting
Zhuang Ma and Zongming Ma. University of Pennsylvania

4:05 PM Individualized Multi-directional Variable Selection
Xiwei Tang and Annie Qu. University of Illinois Urbana and Champaign

4:30 PM Floor Discussion.

Session 17: Statistical methods to advance patient tailoring and precision medicine (Invited)
Room: CONFERENCE ROOM 5C, FIFTH FLOOR
Organizer: Man, Michael; Academia Sinica Taiwan.
Chair: Yu, Danni; Eli Lilly and Company.

3:15 PM Individual Treatment Effect for Better Decision Making in Drug Development and Healthcare
Jia Jia, Qi Tang, Wangang Xie and Richard Rode.

3:40 PM Adaptive threshold design for precision medicine clinical trial
Meijuan Li. Food and Drug Administration

4:05 PM Building a Bayesian decision-theoretic framework to design biomarker-driven studies in early phase c
Danni Yu. Eli Lilly and Company

4:30 PM Development and Application of Shiny Tools for Biomarker Analyses
Qinghua Song. Gilead Sciences, Inc.

4:55 PM Floor Discussion.

Session 18: Challenges in the Analysis of Large Spatial Data (Invited)
Room: CONFERENCE ROOM 5D, FIFTH FLOOR
Organizer: Hu, Juan; Depaul University.
Chair: Wang, Yong; Ohio Northern University.

3:15 PM Construction of Space-Time Matern Correlation Functions
Tonglin Zhang and Hao Zhang. Purdue University

3:40 PM Bootstrap variance estimation for one-per-stratum spatial sampling design
Zhonglei Wang and Zhengyuan Zhu. Iowa State University

4:05 PM The Big Data Issues in Spatial Statistics
Hao Zhang. Purdue University

4:30 PM A Multivariate Spatial Modelling Approach with Nonparametric Cross-covariogram
Yong Wang* and Hao Zhang*. Ohio Northern University Purdue University

4:55 PM Floor Discussion.

Monday, June 26. 8:30AM - 9:45 AM

Opening Ceremony and Keynote Session I (Keynote)
Room: WILLIFORD ROOM
Organizers: ICSA 2017 organizing committee.
Chair: Lanju Zhang.

8:15 AM Welcome
Lanju Zhang, 2017 ICSA Executive Committee Chair, Abbvie Inc

Tony Cai, 2017 ICSA President

8:30 AM Keynote lecture I: Invitation to Improve the Statistical Profession: A Discussion of the Current ASA "Asian Initiative”
Barry D. Nussbaum. American Statistical Association

8:45 AM Keynote lecture II: Personalized Treatment: Sounds heavenly, but where on Earth did they find the right guinea pig for me?
Xiao-Li Meng. Department of Statistics, Harvard University

9:45 AM Floor Discussion.

Monday, June 26. 10:00 AM -11:40 AM

Session 19: New applications in survival and longitudinal data (Invited)
Room: WILLIFORD A, THIRD FLOOR
Organizer: Chen, Yi-Hau; Academia Sinica Taiwan.
Chair: Chaudhuri, Sanjay; National University of Singapore.

10:00 AM Statistical inference on quantile residual life with clustered survival data
Yun Chan Chi, I-Wen Lan and Tsung-Hsien Tsai. National Cheng Kung University
10:25 AM Multi-state Event Analysis with Dynamic Covariates, Time-varying Coefficients and Measurement Errors

*Chuoxin Ma and Jianxin Pan.* School of Mathematics, University of Manchester

10:50 AM Comparison of different approaches for dynamic prediction of survival using longitudinal data

Xuelin Huang. University of Texas MD Anderson Cancer Center

11:15 AM Statistical Monitoring of Clinical Trials with Semi Competing Risks Outcomes

*Toshimitsu Hamasaki¹, Tomoyuki Sugimoto², Scott R. Evans³ and Koko Asakura³.* ¹National Cerebral and Cardiovascular Center ²Kagoshima University ³Harvard T.H. Chan School of Public Health

11:40 AM Floor Discussion.

---

**Session 20:** Recent advances in statistical analysis of genetic/genomic data (Invited)

Room: CONFERENCE ROOM 4J, FOURTH FLOOR

Organizer: Cui, Yuehua; Michigan State University.

Chair: Honglang Wang; Indiana University Purdue University Indianapolis.

10:00 AM Testing variance components in the presence of nuisance boundary parameters

Fei Zou. University of Florida

10:25 AM A group LASSO based approach for GWAS meta-analysis across platforms

*Chi Song¹, Yuan Jiang², Kai Li² and Shuyuan Lou¹.* ¹Ohio State University ²Oregon State University

10:50 AM A semi-parametric test using multiple candidate kernels for gene set association analysis

*Tao He¹, Shaoyu Li², Ping-Shou Zhong³ and Yuehua Cui³.* ¹San Francisco State University ²University of North Carolina at Charlotte ³Michigan State University

11:15 AM Integrating Transcriptional Time Lag Information into Gene Regulatory Network Construction through O

*Yaqun Wang¹, Runze Li² and Rongling Wu².* ¹Rutgers, The State University of New Jersey ²The Pennsylvania State University

11:40 AM Floor Discussion.

---

**Session 21:** Title of session: Bayesian methods in Biostatistics (Invited)

Room: CONFERENCE ROOM 4B, FOURTH FLOOR

Organizer: Daniels, Mike; University of Texas Austin.

Chair: Daniels, Mike; University of Texas Austin.

10:00 AM Bayesian predictive modeling for genomic-based personalized treatment selection

Junsheng Ma¹, Francesca Stingo² and *Brian Hobbs³.* ¹University of Texas, MD Anderson Cancer Center ²University of Forence

10:25 AM Bayesian Methods for Evaluating Air Quality Policies

Corwin Zigler. Harvard TH Chan School of Public Health

10:50 AM A Bayesian Non-Parametric Causal Inference Model for Comparative Effectiveness Research

*Chenguang Wang and Gary Rosner.* Johns Hopkins University

11:15 AM Bayesian Nonparametric Approaches to Causal Inference on Quantiles

*Dandan Xu¹, Micheal Daniels³ and Almut Winterstein².* ¹The University of Texas at Austin ²University of Florida

11:40 AM Floor Discussion.

---

**Session 22:** Causal inference with experimental and observational data (Invited)

Room: CONFERENCE ROOM 4I, FOURTH FLOOR

Organizer: Ding, Peng; UC Berkerley.

Chair: Ding, Peng; UC Berkerley.

10:00 AM Instrumental variables as bias amplifiers with general outcome and confounding

Peng Ding. University of California, Berkeley

10:25 AM Bayesian regression tree models for causal inference

*Richard Hahn¹, Jared Murray² and Carlos Carvalho³.* ¹University of Chicago ²Carnegie Mellon University ³University of Texas

10:50 AM Trustworthy Analysis of Online A/B Tests: Pitfalls, challenges and solutions

Alex Deng, Jiannan Lu and Jonathan Litz. Microsoft Corporation

11:15 AM Using Survival Information in Truncation by Death Problems Without the Monotonicity Assumption

*Fan Yang¹ and Peng Ding².* ¹University of Chicago ²University of California, Berkeley

11:40 AM Floor Discussion.

---

**Session 23:** Recent Developments in Early Phase Dose Finding (Invited)

Room: CONFERENCE ROOM 4H, FOURTH FLOOR

Organizer: Hu, Chih Chi; Sanofi.

Chair: Mao, Xuezhou; Sanofi.

10:00 AM Sequential Quantile Estimation Using Continuous Outcomes: With Applications in Dose Finding

*Chi-Hi Hu¹ and Ying Kuen Cheung².* ¹Sanofi ²Columbia University

10:25 AM Sample size determination for the two-stage continual reassessment method (CRM)

*Cody Chiuzan, Zilan Chai and Ken Cheung.* Columbia University

10:50 AM Parametric Dose Standardization for Two-Agent Phase I-II Trials with Ordinal Efficacy and Toxicity

*Peter Thall.* M.D. Anderson Cancer Center

11:15 AM Discusstant: Ying Yuan

11:40 AM Floor Discussion.
Session 24: Recent Advances in Missing Data Methods (Invited)
Room: CONFERENCE ROOM 4P, FOURTH FLOOR
Organizers: Huang, Chiung-Yu/Hu, Zonghui; Johns Hopkins University; NIH.
Chair: Huang, Chiung-yu; Johns Hopkins University; NIH.

10:00 AM Survival Analysis with Presence of Informative Censoring via Nonparametric Multiple Imputation
\textbullet{} Chengcheng Hu$^1$, Jeremy Taylor$^2$ and Chiu-Hsieh Hsu$^3$.
1University of Arizona 2University of Michigan

10:25 AM Variable selection in the presence of missing data: resampling and imputation
Qi Long$^1$. Changge Chang$^1$ and Brent Johnson$^2$.
1University of Pennsylvania 2University of Rochester

10:50 AM Semiparametric pseudoscore for regression with multidimensional but incompletely observed regressor
\textbullet{} Zonghu Hu, Jing Qin and Dean Pollmann. National Institutes of Health

11:15 AM Limitless regression discontinuity
Adam C Sales$^3$ and Ben Hansen$^2$. 1University of Texas-Austin 2University of Michigan-Ann Arbor

11:40 AM Floor Discussion.

Session 25: Exposure Response Modeling (Invited)
Room: CONFERENCE ROOM 4R, FOURTH FLOOR
Organizer: Huang, Dalong "Patrick"; FDA.
Chair: Qiu, Junshan; FDA.

10:00 AM Statistical Considerations in Exposure Response Modeling
Guoxing Song. FDA

10:25 AM Assay Sensitivity Analysis using Exposure Response Modeling in “Hybrid TQT” Study
\textbullet{} Dalong Huang, Janell Chen, Qianyu Dang and Yi Tsong.
FDA/CDER

10:50 AM Regulatory application of exposure response analyses
Yaning Wang. Food and Drug Administration

11:15 AM A Summary of Dose-Response and Exposure-Response Methodology
\textbullet{} Casey Davis and Bret Musser. Merck & Co.

11:40 AM Floor Discussion.

Session 26: Brain Connectivity and Network Analysis (Invited)
Room: WILLIFORD B, THIRD FLOOR
Organizer: Li, Lexin; UC Berkeley.
Chair: Zhang, Lin; University of Minnesota.

10:00 AM Estimating Brain Pathway Effects Using Large-scale Multilevel Models
\textbullet{} Xi Luo and Yi Zhao. Brown University

10:25 AM Spatially weighted low-rank model for group analysis of functional neuroimaging data
\textbullet{} MiHyee Ahn$^1$, Haipeng Shen$^2$ and Hongtu Zhu$^3$.
1University of Nevada Reno 2University of Hong Kong 3University of Texas MD Anderson Cancer

10:50 AM Estimation of Anatomically Informed Functional Networks via Bayesian Gaussian Graphical Model
Lxavier A. Higgins$^1$, Suprateek Kundu$^2$. 1University of Louisville 2Emory University

Session 27: Advancing Cancer Clinical Trials (Invited)
Room: CONFERENCE ROOM 5G, FIFTH FLOOR
Organizer: Li, Yimei; University of Pennsylvania.
Chair: Hwang, Wei-Ting; University of Pennsylvania.

10:00 AM Suspension of accrual in phase II cancer clinical trials
\textbullet{} Yimei Li$^1$, Rosemarie Mick$^2$ and Daniel Heitjan$^3$.
1University of Pennsylvania 2Southern Methodist University

10:25 AM Dose finding designs for late-onset and cumulative toxicities
\textbullet{} Shing Lee$^1$, Moreno Urso$^2$, Ying Kaen Cheung$^3$ and Sarah Zohar$^4$.
1Columbia University 2INSERM

10:50 AM A multi-stage adaptive enrichment design to improve efficiency of early phase clinical trials
Brandon Luber and Hao Wang. Johns Hopkins University

11:15 AM Design considerations in immunotherapy trials with multiple targets and indications
Kay See Tan. Memorial Sloan Kettering Cancer Center

11:40 AM Floor Discussion.

Session 28: Recent Advance in Platform Clinical Trials in Era of Precision Medicine (Invited)
Room: WILLIFORD C, THIRD FLOOR
Organizer: Lin, Jianchang; Takeda.
Chair: Yang, Fang; Vertex Pharmaceuticals.

10:00 AM A Subgroup Cluster Based Bayesian Adaptive Design for Precision Medicine
\textbullet{} Wentian Guo$^1$, Yuan Ji$^2$ and Daniel Catenacci$^2$.
1Laiya Consulting, Inc. 2University of Chicago

10:25 AM Increasing Efficiency of Oncology Basket Trials Using a Bayesian Approach
\textbullet{} Rong Liu. University of Toledo

10:50 AM Early Signal Detection in Basket Trial via Bayesian Approach
\textbullet{} Charlie Cao and Wenwen Zhang. Takeda

11:15 AM Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials
\textbullet{} Jianchang Lin$^1$ and Veronica Bunn$^2$.
1University of Texas 2Florida State University

11:40 AM Floor Discussion.

Session 29: Aspects of Statistical Inference (Invited)
Room: CONFERENCE ROOM 5H, FIFTH FLOOR
Organizer: Markatou, Marianne; University of Buffalo.
Chair: Markatou, Marianne; University of Buffalo.

10:50 AM Estimation of Anatomically Informed Functional Networks via Bayesian Gaussian Graphical Model
Lxavier A. Higgins$^1$, Suprateek Kundu$^2$. 1University of Louisville 2Emory University

11:15 AM Sparse multivariate multiple regression to identify reference ability specific resting state network
\textbullet{} Seonjoo Lee$^1$, Christian Habeck$^2$ and Yaakov Stern$^2$.
1Department of Biostatistics, Columbia University 2Department of Neurology, Columbia University

11:40 AM Floor Discussion.
Session 30: Robustness Aspects of Optimal Experimental Design (Invited)
Room: CONFERENCE ROOM 5F, FIFTH FLOOR
Organizer: O'Brien Timothy E.; Loyola University Chicago.
Chair: O'Brien Timothy E.; Loyola University Chicago.
10:00 AM Developments for Information-Based Subdata Selection
John Stufken. Arizona State University
10:25 AM Robustness of Block Designs to Data Loss
JP Morgan. Department of Statistics, Virginia Tech
10:50 AM EW Optimality for Robust Experimental Design under Generalized Linear Models
Jie Yang. University of Illinois at Chicago
11:15 AM Flexible Optimal Design Strategies
Timothy O'Brien. Loyola University Chicago
11:40 AM Floor Discussion.

Session 31: Opportunities to innovatively leverage historical information to improve medical product development (Invited)
Room: BOULEVARD AB ROOM, SECOND FLOOR
Organizer: Price, Karen; Eli Lilly and Company.
Chair: Brenda Crowe; Eli Lilly and Company.
10:00 AM Leveraging historical information in clinical programs
Freda Cooner. Sanofi
10:25 AM Bayesian methods toward a feasible and informative clinical trial in small populations
Margaret Gamalo. Eli Lilly & Co
10:50 AM Bayesian Historical Borrowing Method with Case Study in a Phase 3 HCV Trial
Ran Liu1, Qi Tang2, Martin King3, Bo Fu1, Sandra Lovell1 and Alan Hartford1. 1AbbVie Inc. 2Sanofi
11:15 AM Discussant: John Scott
11:40 AM Floor Discussion.

Session 32: Student Session (I) (Student Winner Session)
Room: CONFERENCE ROOM 5I, FIFTH FLOOR
Chair: Xuan Zhou; University of North Carolina.
10:00 AM Sequential Outcome-Weighted Multicategory Learning for Estimating Optimal Individualized Treatment
Xuan Zhou1, Yuanjia Wang2 and Donglin Zeng3. 1University of North Carolina at Chapel Hill 2Columbia University
10:25 AM Fused Gaussian Process for Very Large Spatial Data
Palong Ma and Emily Kang. University of Cincinnati
10:50 AM Matrix Completion with Covariate Information
Xiaojun Mao, Songxi Chen and Raymond Wong. Iowa State University
11:15 AM Floor Discussion.

Session 33: Multiple tests in Clinical trials (Invited)
Room: CONFERENCE ROOM 5E, FIFTH FLOOR
Organizer: Quan, Hui; Sanofi.
Chair: Luo, Xiaodong; Sanofi.
10:00 AM Bonferroni - based gatekeeping procedure with retesting option
Zhijing Qiu1, Wenge Guo2 and Sanat Sarkar3. 1Biostatistics and Programming, Sanofi US 2New Jersey Institute of Technology 3Temple University
10:25 AM A multi-dimensional multiplicity problem in a clinical trial
Meehyung Cho and Zhijing Qiu. Sanofi
10:50 AM Graphical Approaches for Multiplicity Adjustment in Clinical Trials
Zhi Wen. Sanofi
11:15 AM A seamless phase II/III/IV clinical trial design with different endpoints for different phases
Hui Quan, Tianyue Zhou and Peng-Liang Zhao. Sanofi
11:40 AM Floor Discussion.

Session 34: Novel methods and practical considerations in missing data analysis (Invited)
Room: CONFERENCE ROOM 5A, FIFTH FLOOR
Organizer: Mao, Xuezhou; Sanofi.
Chair: Hu, Chih-Chi; Sanofi.
10:00 AM Tipping point sensitivity analysis for incomplete event count data
Xuezhou Mao1, Li Qi1, Hui Quan1, Mei Zhang2, Lin Wang1 and Lynn Wei1. 1Biostatistics and Programming, Sanofi U.S. 2Translational Medicine, Sanofi U.S.
10:25 AM Options for implementing pattern-mixture-based sensitivity analyses
Ilya Lipkovich, Bohdana Ratitch and Michael O’Kelly. QuintilesIMS
10:50 AM Diagnosing Missing Always at Random in Multivariate Data
Ivor Bojinov, Natech Pillai and Donald Rubin. Harvard Statistics Department
Session 35: recent developments in parsimonious modelling (Invited)

Room: CONFERENCE ROOM 4Q, FOURTH FLOOR
Organizer: Li, Gen; Columbia University.
Chair: Li, Gen; Columbia University.

10:00 AM On the analysis of Bregman-surrogate algorithms for non-convex optimization
   ▪ Zhifeng Wang and Yiyouan She. Florida State University

10:25 AM Incorporating Covariates into Integrated Factor Analysis of Multi-View Data
   Gen Li1 and Sangkyu Jung2. 1Columbia University 2University of Pittsburgh

10:50 AM Embracing the Blessing of Dimensionality in Factor Models
   ▪ Quefeng Li1, Guang Cheng2, Jianqi Fan3 and Yuyan Wang3. 1UNC Chapel Hill 2Purdue University 3Princeton University

11:15 AM An efficient approach to discriminant analysis on tensor data with covariates
   ▪ Yaqing Pan, Qing Mai and Xin Zhang. Department of Statistics, Florida State University

11:40 AM Floor Discussion.

Session 36: Recent Advances in Enrichment Strategies: Methods and Case Studies (Invited)

Room: CONFERENCE ROOM 5B, FIFTH FLOOR
Organizer: Tang, Qi; Sanofi.
Chair: Tang, Qi; Sanofi.

10:00 AM Bayesian Random Partition Models for Subgroup Identification
   Wentian Guo1 and Yuan Ji2. 1Laiya Consulting, Inc. 2NorthShore U. HealthSys/The University of Chicago

10:25 AM Optimization of Multi-arm Adaptive Enrichment Designs via Simulated Annealing
   ▪ Joshua Betz, Michael Rosenblum, Jon Steingrimson and Adi Gherman. Johns Hopkins Bloomberg School of Public Health

10:50 AM On Group Sequential Enrichment Design for Basket Trial
   ▪ Shuai Yuan1, Aiyiing Chen2 and Li He3. 1Merck 2Sanofi Pasteur

11:15 AM Case studies of Enrichment Strategies in Oncology
   Ming Zhu. Sanofi Pasteur

11:40 AM Floor Discussion.

Session 37: Recent robust statistical methodology in data analysis (Invited)
Room: CONFERENCE ROOM 5E, FIFTH FLOOR
Organizer: Chen, Yi-Hau; Academia Sinica Taiwan.
Chair: Chen, Hua Yun; University of Illinois Chicago.

1:00 PM Robust likelihood inference for two correlated multinomial distributions
   Tsung-Shan Tsou. Institute of Statistics, National Central University, Jhongli, Taiwan

1:25 PM Hypothesis testing in functional linear models
   ▪ Chongzhi Di, Yu-Ra Su and Li Hsu. Fred Hutchinson Cancer Research Center

1:50 PM Drug cumulative effect in observational study with application on ADHD patients.
   ▪ Shinsheng Yuan and Kang-chung Yang. Institute of Statistical Sciences, Academia Sinica

2:15 PM Floor Discussion.

Session 38: Modern Developments in Statistical Classification and Selection (Invited)
Room: WILLIFORD A, THIRD FLOOR
Organizer: Han, Xu; Temple University.
Chair: Tong, Xin; University of Southern California.

1:00 PM Model-free variable selection for the regression mean
   Yaexiao Dong. Temple University

1:25 PM High-dimensional interaction selection via regularization
   Ning Hao1, Yang Feng2 and Helen Zhang3. 1University of Arizona 2Columbia University 3UCLA

1:50 PM An Umbrella Algorithm for Neyman-Pearson Classification
   ▪ Xin Tong1, Yang Feng2 and Jingyi Li3. 1University of Southern California 2Columbia University 3UCLA

2:15 PM Model-based Clustering for Dynamic Networks
   ▪ Kevin Lee, Lingzhou Xue and David Hunter. The Pennsylvania State University

2:40 PM Floor Discussion.

Session 39: Non- and semi-parametric models in the large-scale genomic data analyses (Invited)
Room: CONFERENCE ROOM 4J, FOURTH FLOOR
Organizer: He, Tao; San Francisco State University.
Chair: He, Tao; San Francisco State University.

1:00 PM Improved shrunken centroid classification for better variable selection in genomics data
   Yukun Li, Haiyan Wang and Huaiyu Zhang. Kansas State University

1:25 PM A multivariate semiparametric approach for gene-environment interactions
   ▪ Cen Wu1, Yuehua Cui2 and Shuangge Ma3. 1Department of Statistics, Kansas State University 2Dept. of Stat. and Prob. Michigan State University 3Department of Biostatistics, Yale University

2017 ICSA Applied Statistics Symposium, Chicago, June 25-28
Session 40: Integrating metagenomics and other -omics data

1:00 PM Mediation Analysis in Microbiome Studies
   Hongzhe Li. University of Pennsylvania

1:25 PM An Optimum Microbial Association Test
   Hyunwook Koh, Martin Blaser and Hui Lin. New York University School of Medicine

1:50 PM Integrative Analysis of High-dimensional Microbiome and Host Genomic Data
   Lingling An and Meng Lu. University of Arizona

2:15 PM Floor Discussion.

Session 41: Design and Uncertainty Quantification for Computer Experiments (Invited)

1:00 PM A Bayesian Approach to Model Selection & Uncertainty Quantification in Brain Injury Simulations
   Sandeep Madireddy1, Kumar Vemgantii and Emily Kang2. 1Argonne National Laboratory 2University of Cincinnati

1:25 PM Learning About Physical Parameters - The Importance of Model Discrepancy
   Jenny Brynjarsdottir1 and Anthony OHagan2. 1Case Western Reserve University 2University of Sheffield

1:50 PM A General Construction for Space-filling Latin Hypercubes
   Chunyang Lin1 and Lulu Kang2. 1Queen’s University, Canada 2Illinois Institute of Technology

2:15 PM Sequential design of experiment for multiple computer models
   Bledar Konomi. University of Cincinnati

2:40 PM Floor Discussion.

Session 42: Statistical Methods for Network Data (Invited)

1:00 PM A new SVD approach to optimal topic estimation
   Tracy Ke and Minze Wang. University of Chicago

1:25 PM Regression analysis of networked data
   Yan Zhou1 and Peter Song2. 1University of Michigan

1:50 PM Time-varying network estimation from high-dimensional time series
   Mengyu Xu1, Xiaohui Chen3 and Wei Biao Wu2. 1University of Central Florida 2University of Illinois Urbana-Champaign 3University of Chicago

2:15 PM Testing and Estimation of Social Network Dependence with Time to Event Data
   Rui Song. North Carolina State University

2:40 PM Floor Discussion.

Session 43: Selected Topics in Designing and Conducting Safety Studies (Invited)

1:00 PM Post-market evaluation of drug safety
   John Yap1, Brenda Crowe2, Simone Pinheiro1 and Yi Yvonne Huang3. 1Food and Drug Administration 2Eli Lilly and Company 3University of Maryland

1:25 PM Safety Signal Detection when Incidence Percentages are Biased
   Brenda Crowe. Eli Lilly and Company

1:50 PM Meta-analysis of observational studies for drug safety
   John Yap. FDA

2:15 PM Latent Propensity Score Approach Allowing Covariate Measurement Error in Observational Studies
   Yi Huang1, Elande Baro2, Andrew Raim3, Anindya Roy2, Dennis Cox4, Karen Bandeen-Roche2, and Yi Huang3. 1US Census Bureau 2US Census Bureau 3University of Maryland 4US Census Bureau 4US Census Bureau 5Eli Lilly and Company

2:40 PM Floor Discussion.

Session 44: Advanced Bayesian Methods with Applications to Medical Data (Invited)

1:00 PM Bayesian semiparametric analysis of mixed effects models with applications to dose-response studies
   Taeryon Choi. Korea University

1:25 PM Flexible Priors for Covariance Functions
   Dennis Cox. Rice University

1:50 PM Bayesian region selection in functional regression
   Hongxiao Zhu and Yizhi Sun. Virginia Tech

2:15 PM Floor Discussion.

Session 45: Statistical and Regulatory Considerations and Opportunities for Rare Diseases Clinical Trials (Invited)

1:00 PM Opportunities for Rare Diseases Clinical Trials
   Pei Geng, Xiaoran Tong and Qing Lu. Michigan State University

1:25 PM Bayesian region selection in functional regression
   Rui Song. North Carolina State University

1:50 PM Metabolic phenotypes of rare disease models
   John Yap. FDA
Monday, June 26. 1:00 PM - 2:40 PM

**Session 46: Adaptive and Innovative Study Designs in Clinical Trials (Invited)**
Room: CONFERENCE ROOM 5H, FIFTH FLOOR
Organizer: Xu, Tu; AbbVie Inc.
Chair: Liu, Ran; AbbVie Inc.

1:00 PM  
Biomarker-driven clinical trial design in precision medicine  
Jing Wang. Pfizer, Inc.

1:25 PM  
New CARA Designs for Personalized Medicine and Their Statistical Inference  
Wanying Zhao and Feifang Hu. The George Washington University

1:50 PM  
Seamless Phase 2/3 Study Design with an Oncology Example  
Zhaoyang Teng, Guohui Liu and Yi Liu. Takeda Pharmaceuticals

2:15 PM  
Optimal flexible sample size design with interim dose determination  
Lu Cui and Lanju Zhang. AbbVie Inc.

2:40 PM  
Floor Discussion.

**Session 47: New machine learning tools for complex data (Invited)**
Room: WILLIFORD B, THIRD FLOOR
Organizer: Yu, Guan; University of Buffalo.
Chair: Xie, Yuying; Michigan State University.

1:00 PM  
Sparse Tensor Response Regression  
Will Wei Sun and Lexin Li. University of Miami

1:25 PM  
GENERAL FRAMEWORK FOR ASSOCIATION ANALYSIS OF HETEROGENEOUS DATA  
Gen Li and Irina Gaynanova. Columbia University. Department of Biostatistics

1:50 PM  
Hierarchical models for predicting above ground biomass with 3D LiDAR signals  
Yazhen Zhou, Andrew Finley and Bruce Cook. University of Nebraska Lincoln

2:15 PM  
Optimal Sparse Linear Prediction for Block-missing Multimodality Data without Imputation  
Guan Yu, Quefeng Li, Dinggang Shen and Yafeng Liu. State University of New York at Buffalo

2:40 PM  
Floor Discussion.

**Session 48: Statistical methods for biomarker data analysis (Invited)**
Room: CONFERENCE ROOM 5F, FIFTH FLOOR
Organizer: Zhang, Donghui; Sanofi.
Chair: Zhang, Donghui; Sanofi.

1:00 PM  
Systematic Evaluation of Statistical Methods in Subgroup Analysis  
Xiaojing Wang and Yang Liu. University of Connecticut

1:25 PM  
Quantitative Reproducibility Analysis for Identifying Reproducible Targets from High-Throughput  
Wenfei Zhang. Sanofi

1:50 PM  
On Pharmacodynamic Biomarker for Early Decision Making  
Atalanta Ghosh. Janssen Research and Development

2:15 PM  
Strategies for Clinical Development of Predictive Biomarkers  
Glen Laird. Vertex Pharmaceuticals

2:40 PM  
Floor Discussion.

**Session 49: Big Data and Healthcare Analytics (Invited)**
Room: BOULEVARD AB ROOM, SECOND FLOOR
Organizer: Zhang, Lingsong; Purdue University.
Chair: Qiao, Xingye; Binghamton University.

1:00 PM  
Integrative Cox Regression for Modeling Uncertain Survival Records due to Imperfect Record Linkage  
Wenjie Wang, Robert Aseltine and Jun Yan. University of Connecticut

1:25 PM  
Pragmatic Clinical Trials in Drug Development  
Dingfeng Jiang. AbbVie

1:50 PM  
Big Data and Healthcare in Real World: Challenges and Opportunities  
Yifan Xu. IBM Watson Health

2:15 PM  
Angle based Multicategory Distance-weighted SVM  
HUI SUN, Bruce Craig and Lingsong Zhang. Purdue University

2:40 PM  
Floor Discussion.

**Session 50: High-dimensional and Complex Data (Invited)**
Room: CONFERENCE ROOM 5I, FIFTH FLOOR
Organizer: Zhang, Qingyang; University of Arkansas.
Chair: Zhang, Qingyang; University of Arkansas.

1:00 PM  
Whiteout: Gaussian Adaptive Regularization Noise in Deep Neural Networks  
Fang Liu, Yinan Li and Ruoyi Xu. University of Notre Dame

1:25 PM  
High-dimensional robust regression  
Po-Ling Loh. UW-Madison
1:50 PM Analyzing tree-based survival models: splitting rule, adaptive concentration and consistency
Yifan Cui\textsuperscript{1}, Ruoqing Zhu\textsuperscript{3}, Mai Zhou\textsuperscript{3} and Michael Kosorok\textsuperscript{2}. \textsuperscript{1}University of North Carolina at Chapel Hill \textsuperscript{2}University of Illinois Urbana-Champaign \textsuperscript{3}University of Kentucky

Session 51: The state of some statistical challenges in Pharma 10 years hence (Invited)
Room: WILLIFORD C, THIRD FLOOR
Organizer: Cicconetti, Greg; AbbVie Inc.
Chair: Cicconetti, Greg; AbbVie Inc.

1:00 PM The state of some statistical challenges in Pharma 10 years hence
\textbullet\ Walt Offen\textsuperscript{1}, Frank Rockhold\textsuperscript{2} and Qi Jiang\textsuperscript{3}. \textsuperscript{1}AbbVie Inc \textsuperscript{2}Duke University \textsuperscript{3}Amgen

1:30 PM Panel Discussion: Qi Jiang, Frank Rockhold, Walt Offen

2:40 PM Floor Discussion.

Session 52: New Development in biomedical applications (Invited)
Room: CONFERENCE ROOM 5A, FIFTH FLOOR
Organizer: Mei-Ling Ting Lee; University of Maryland.
Chair: Chengsheng Jiang; University of Maryland.

1:00 PM A New Measure of Synchronization to Quantify Brain Connectivity
Yang Zhou\textsuperscript{1}, Jane-Ling Wang\textsuperscript{1}, Hans-Georg Mueller\textsuperscript{1} and Owen Carmichael\textsuperscript{2}. \textsuperscript{1}University of California, Davis \textsuperscript{2}Pennington Biomedical Research Center

1:25 PM On Exact and Approximate Distributions of K-homopolymer for iid and Markov Dependent DNA Sequences
James Fa\textsuperscript{1}, Hsing-Ming Chang\textsuperscript{2} and Wendy Lou\textsuperscript{3}. \textsuperscript{1}University of Manitoba \textsuperscript{2}National Cheng Kung University \textsuperscript{3}University of Toronto

1:50 PM Statistical Analysis of Single Cell DNA methylation Data
Yongseok Park and Zhiguang Huo. University of Pittsburgh

2:15 PM A Novel Statistical Model for Detecting DNA Methylation Marks Collected in Paired Design
Yanjing Li\textsuperscript{1}, Jarrett Morrow\textsuperscript{2}, Benjamin Raby\textsuperscript{2}, Kelan Tantisira\textsuperscript{2}, Scott Weiss\textsuperscript{2}, Wei Huang\textsuperscript{3} and Wei Liang Qiu\textsuperscript{2}. \textsuperscript{1}Zhejiang University \textsuperscript{2}Harvard Medical School

2:40 PM Floor Discussion.

Session 53: Utilization of historical data into clinical trial development (Invited)
Room: CONFERENCE ROOM 4Q, FOURTH FLOOR
Organizer: Liu, Jingyi; Eli Lilly and Company.
Chair: Zhang, Xiang ; Eli Lilly and Company.

1:00 PM The utility of collaboration: the TransCelerate Placebo/Standard of Care database
Jessica Lim. GSK

1:25 PM A Bayesian Approach to Incorporate Historical Information in Clinical Trials
\textbullet\ Judy Li\textsuperscript{1}, Wei-chen Chen\textsuperscript{2} and John Scott\textsuperscript{2}. \textsuperscript{1}Regeneron, \textsuperscript{2}FDA

1:50 PM Leveraging Historical Control Data: The Bayesian Augmented Control Method, with Case Studies and R Implementation
\textbullet\ Hongtiao Zhang\textsuperscript{1}, Saurabh Mukhopadhyay\textsuperscript{1}, Qi Tang\textsuperscript{2}, Ran Liu\textsuperscript{2}, Kun Chen\textsuperscript{2}, Martin King\textsuperscript{1} and Bo Fu\textsuperscript{1}. \textsuperscript{1}AbbVie Inc. \textsuperscript{2}Sanofi

2:15 PM Discussant: Kert Viele, Berry Consultants

2:40 PM Floor Discussion.

Session 54: Recent advances in statistical genomics in personalized medicine (Invited)
Room: CONFERENCE ROOM 5B, FIFTH FLOOR
Organizer: Huang, Bo; Pfizer.
Chair: Tang, Rui (Sammi); Vertex Pharmaceuticals.

1:00 PM Integrating data-driven priors into GWAS and mediation analysis of GWAS with penalized regression
Sanaz Keles. University of Wisconsin, Madison

1:25 PM Statistical method for improving efficiency of CRISPR sgRNA design
\textbullet\ Pei Fen Kuan\textsuperscript{1}, Scott Powers\textsuperscript{1}, Shuyao He\textsuperscript{1}, Kaiqiao Li\textsuperscript{1}.
Xiaoyu Zhao\textsuperscript{3} and Bo Huang\textsuperscript{2}. \textsuperscript{1}Stony Brook University \textsuperscript{2}Pfizer Inc.

1:50 PM Unbiased estimation of parent-of-origin effects using RNA-seq data from human
Vasyl Zhabotynsky\textsuperscript{1} and \textbullet\ Wei Sun\textsuperscript{2}. \textsuperscript{1}UNC Chapel Hill \textsuperscript{2}Fred Hutchinson Cancer Research Center

2:15 PM Sparse additive index model for survival prediction with genomic data
Muxuan Liang, Kam-Wah Tsui and \textbullet\ Sijian Wang. University of Wisconsin-Madison

2:40 PM Floor Discussion.

Monday, June 26. 2:50 PM - 4:30 PM

Session 55: New developments in high-dimensional data analysis (Invited)
Room: CONFERENCE ROOM 5A, FIFTH FLOOR
Organizer: Chen, Yi-Hau; Academia Sinica Taiwan.
Chair: Huang, Xuelin; University of Texas MD Anderson.

2:50 PM An Inverse-regression Method of Dependent Variable Transformation for Dimension Reduction with Non-linear Confounding
Heng-Hui Lue. Tunghai University

3:15 PM Dynamic prediction involving high-dimensional factors based on the joint frailty-copula model
\textbullet\ Takeshi Emura\textsuperscript{1}, Masahiro Nakatocchi\textsuperscript{2}, Shigeyuki Matsui\textsuperscript{3}.
Hirofumi Michimae\textsuperscript{4} and Virginie Rondeau\textsuperscript{5}. \textsuperscript{1}National...
Session 56: The Power of Integrative Genomics Data Analysis and Related Statistical Issues (Invited)
Room: CONFERENCE ROOM 4F, FOURTH FLOOR
Organizer: Li, Hongzhe; University of Pennsylvania.
Chair: Tang, Zhengzheng; Vanderbilt University.

2:50 PM A new algorithm for simultaneous clustering of genes in two species based on homology and expression
Yidan Sun and Jingxi Jessica Li. UNIVERSITY OF CALIFORNIA, LOS ANGELES

3:15 PM A statistical framework of mapping risk genes from de novo mutations in whole-genome sequencing study
Xin He1, Yuwen Liu1, Ercument Cicek2, Yanyu Liu1, James Noonan3 and Zhongsheng Sun4. 1University of Chicago 2Bilkent University, Turkey 3Yale University 4Chinese Academy of Science

4:05 PM Practical experience in integrative genomics. 
Yuanhong Sun1, Yanyu Liu1, Wei Pan1, Youyao Xie1, Nianming Fu1,2, Yu Zhang1, Feng Gao1,2,3, An Qian1,2, Yu Zhang1,2, Jichuan Xie1. 1Shenzhen Institutes of Advanced Technology, 2University of Chinese Academy of Science, 3University of Massachusetts Amherst.

Session 57: The Art of Evaluating the Surrogate Outcomes (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Liu, Lan; University of Minnesota.
Chair: Liu, Lan; University of Minnesota.

2:50 PM Estimating Treatment Effects Using Multiple Surrogates: The Role of the Surrogate Score and Index
Hyunseung Kang. University of Wisconsin Madison

3:15 PM Estimation of the Optimal Surrogate Using SuperLearner and TMLE
Brenda Price1, Peter Gilbert2 and Mark van der Laan3. 1University of Washington 2Fred Hutchinson Cancer Research Center 3University of California, Berkeley

3:40 PM Novel Criteria to Exclude the Surrogate Paradox and Their Qualities
Yunqian Yin1, Lan Liu2, Zhi Geng1 and Peng Luo3. 1Peking University 2University of Minnesota 3Shenzhen University

4:05 PM Floor Discussion.

Session 58: New Development on Matching Methodology for Causal Inference (Invited)
Room: CONFERENCE ROOM 4I, FOURTH FLOOR
Organizer: Lu, Bo; Ohio State University.
Chair: Song, Chi; Ohio State University.

2:50 PM Strong control of the familywise error rate in observational studies discovering effect modification
Jesse Hsu1, Jose Zubizarreta2, Dylan Small3 and Paul Rosenbaum1. 1University of Pennsylvania 2Columbia University

3:15 PM Propensity Score Matching for Clustered Data
Mi-Ok Kim1, Bo Lu2 and Chunyan Liu3. 1University of California, San Francisco 2Ohio State University 3Cincinnati Children’s Hospital Medical Center

3:40 PM Near-optimal matching algorithm for multi-group observational studies
Bo Lu. The Ohio State University

4:05 PM Optimal Multilevel Matching Using Network Flows: An Application to a Summer Reading Intervention
Samuel Pimentel1, Lindsay Page2, Matthew Lenard3 and Lake Keele4. 1University of Pennsylvania 2University of Pittsburgh 3Wake County Public Schools 4Georgetown University

4:30 PM Floor Discussion.

Session 59: Recent Advances in Spatial Statistics: Theory and Application (Invited)
Room: CONFERENCE ROOM 4H, FOURTH FLOOR
Organizer: Zhu, Zhengyuan; Iowa State University.
Chair: Zhu, Zhengyuan; Iowa State University.

2:50 PM Inference for spatial autocorrelation on a stream network
Dale Zimmerman. University of Iowa

3:15 PM Multivariate parametric models in ecology, with focus on addressing spatiotemporal variation
Brian Gray1 and Karl Oskar Ekvall2. 1US Geological Survey 2University of Minnesota

3:40 PM Variogram models on spheres of all dimensions
Xintong Li1, Juan Du1 and Chunsheng Ma2. 1Kansas State University 2Wichita State University

4:05 PM Flexible and efficient estimating equations for variogram estimation
Xiaohui Chang1, Ying Sun2 and Yongtao Gao3. 1Oregon State University 2King Abdullah University of Science and Technology 3University of Miami

4:30 PM Floor Discussion.

Session 60: Large-Scale Statistical Inference (Invited)
Room: WILLFORD B, THIRD FLOOR
Organizer: Cai, Tony; University of Pennsylvania.
Chair: Cai, Tony; University of Pennsylvania.

2:50 PM On Polynomial Time Methods for Exact Low Rank Tensor Completion
Dong Xia and Ming Yuan. University of Wisconsin

3:15 PM Statistical and Computational Guarantees of Lloyd’s Algorithm and Its Variants
Yu Lu and Harrison Zhou. Yale University
3:40 PM A General Framework for Information Pooling in Two-Sample Sparse Inference
   *Yin Xia*\(^1\), *Tony Cai*\(^2\) and *Wenguang Sun*\(^3\). \(^1\)Fudan University \(^2\)University of Pennsylvania \(^3\)University of Southern California

4:05 PM Floor Discussion.

**Session 61: Big Data and Interactions (Invited)**
Room: CONFERENCE ROOM 4F, FOURTH FLOOR
Organizer: Zhao, Linda; University of Pennsylvania.
Chair: Zhao, Linda; University of Pennsylvania.

2:50 PM TextM for Crowdsourcing in Epidemiology: Interface of Statistics and Computer Science
   *Jiayang Sun*\(^1\), *Katie Pezzott*\(^2\), *Lu Wang*\(^3\) and *Rebecca Carter*\(^4\). \(^1\)Case Western Reserve University \(^2\)Battelle Memorial Institute

3:15 PM Hierarchical Sparse Modeling of Interactions
   *Xiaohan Yan*\(^5\) and *Jacob Bien*\(^6\). \(^5\)Cornell University \(^6\)University of Southern California

3:40 PM Boosting Gene Mapping Power and Efficiency with Efficient Exact Variance Component Tests of SNP Sets
   *Jin Zhou*\(^1\), *Dandi Qiao*\(^2\), *Michael Cho*\(^3\) and *Hua Zhou*\(^4\). \(^1\)University of Arizona \(^2\)Harvard School of Public Health \(^3\)Harvard Medical School \(^4\)University of California, Los Angeles

4:05 PM Gene-based segregation method for identifying rare variants in family-based sequencing studies
   *Dandi Qiao*\(^1\), *Christoph Lange*\(^2\), *Nan Laird*\(^3\), *Sangho Won*\(^4\), *Craig Hersh*\(^5\), *Jarrett Morrow*\(^6\) and *Brian Hobbs*\(^7\). \(^1\)Brigham and Women’s Hospital \(^2\)Harvard T.H. Chan School of Public Health \(^3\)Harvard Medical School

4:30 PM Floor Discussion.

**Session 62: Recent Advances in Assessment of Agreement for Clinical and Lab Data (Invited)**
Room: CONFERENCE ROOM 4R, FOURTH FLOOR
Organizer: Pan, Yi; CDC.
Chair: Ding, Ying; University of Pittsburgh.

2:50 PM Bayesian Estimate of Concordance Correlation Coefficient
   *Dai Feng*. Merck & Co., Inc

3:15 PM Methods for assessing the reliability of quality of life based on SF-36
   *Yi Pan*. Centers for Disease Control and Prevention

3:40 PM Quantifying an Agreement Study
   *Jason Liao and Jialin Xu*. Merck

4:05 PM Discussant: Ruosha Li

4:30 PM Floor Discussion.

**Session 63: Integrating big and complex data with new statistical tools (Invited)**
Room: CONFERENCE ROOM 5G, FIFTH FLOOR
Organizer: Kong, Linglong; University of Alberta.
Chair: Kong, Linglong; University of Alberta.

2:50 PM SPATIAL-TEMPORAL GAUSSIAN STATE-SPACE MODELING
   *Debashis Mandal and Chunxiao Wang*. Oregon State University

3:15 PM Accounting for uncertainty in smoothing individual longitudinal profiles
   *Joel Dubin and Shoa Chenouri*. University of Waterloo

3:40 PM Shape Constrained Inference in LASSO Regularized Regression
   *Matus Maciak*\(^1\), *Ivan Mizera*\(^2\) and *Gabriela Ciuperca*\(^3\). \(^1\)Charles University \(^2\)University of Alberta \(^3\)University Lyon

4:05 PM Discussant: Dengdeng Yu

4:30 PM Floor Discussion.
Session 66: Recent advances in non-standard survival data
(Invited)
Room: CONFERENCE ROOM 5F, FIFTH FLOOR
Organizer: Chan, Kwun Chuen Gary; University of Washington.
Chair: Di, Chongzhi; Fred Hutchinson Cancer Research Center.
2:50 PM Current Status Data in the Presence of A Terminal Event
Lu Mao. University of Wisconsin-Madison
3:15 PM Outcome-dependent sampling with interval-censored failure
time data
• Qingning Zhou, Jianwen Cai and Haibo Zhou. University of North Carolina at Chapel Hill
3:40 PM Semiparametric Estimation of the Scale-Change Model with
Panel Count Data under Informative Exams
Sy Han Chiou¹. • Gongjun Xu², Jun Yan³ and Chiung-Yu
Huang⁴. ¹Harvard University ²University of Michigan
³University of Connecticut ⁴Johns Hopkins University
4:05 PM Semiparametric regression models for current duration data
Kwun Chuen Gary Chan. University of Washington
4:30 PM Floor Discussion.

Session 67: Recent advance in bioinformatics and computational
biology (Invited)
Room: BOULEVARD AB ROOM, SECOND FLOOR
Organizer: Wang, Zuoheng; Yale University.
Chair: Wang, Zuoheng; Yale University.
2:50 PM Personalized Risk Prediction for Glaucoma
X. Raymond Gao. University of Illinois at Chicago
3:15 PM Integrating Diverse Genomic Data to Estimate Multiple Networks
Yuping Zhang. University of Connecticut
3:40 PM A flexible R package to identify high quality gene modules from
complex Omic networks
Xinyu Zhang¹. • Ying Hu², Chanhua Yan² and Ke Xu¹.
¹Department of Psychiatry, Yale School of Medicine
²National Cancer Institute
4:05 PM An Isoform-free Model for Differential Expression Analysis
in Large Sample RNA-seq Data
Yang Liu. Abbvie
4:30 PM Floor Discussion.

Session 68: Recent advances in statistical genetics (Invited)
Room: CONFERENCE ROOM 5I, FIFTH FLOOR
Organizer: Wu, Xiaowei; Virginia Tech.
Chair: Wu, Xiaowei; Virginia Tech.
2:50 PM A powerful framework for integrating eQTL and GWAS
summary data
Zhiyuan Xu¹. • Chong Wu¹, Peng Wei² and Wei Pan¹.
¹University of Minnesota ²MD Anderson Cancer Research Center
3:15 PM A novel test by testing an optimally weighted combination
of variants with multiple traits
Zhenghuan Wang, • Quying Sha, Kai Zhang and Shuanglin
Zhang. Michigan Technological University
3:40 PM Dictionary learning based genotype imputation to improve
power for association testing
• Xiaowei Wu and Hongxiao Zhu. Virginia Tech
4:05 PM Floor Discussion.

Session 69: Urging a paradigm change: BFF inferences in the
era of data science (Invited)
Room: CONFERENCE ROOM 5E, FIFTH FLOOR
Organizer: Xie, Minge; Rutgers University.
Chair: Paul Edlefsen; Fred Hutchinson Cancer Research Center.
2:50 PM On the cross correlations under high dimension
Han Xiao. Rutgers University
3:15 PM Non-penalized variable selection in high-dimensional linear
model via generalized fiducial inference
Jonathan Williams and • Jan Hannig. University of North Carolina at Chapel Hill
3:40 PM Higher-Order BFF
Todd Kuffner. Washington University in St. Louis
4:05 PM An Objective Prior for Hyperparameters in Normal Hierarchical Models
James Berger¹. • Dongchu Sun² and Chengyu Song³.
¹Duke University ²University of Missouri ³East China Normal University
4:30 PM Floor Discussion.

Session 70: Panel Session on Leadership (Top Contributed)
Room: WILLIFORD A, THIRD FLOOR
Organizer: Lei Wang; Lotus Group.
Chair: Helena Fan; Lotus Group.
2:50 PM First and Foremost... Know Thyself
Chito Hernandez. Biomarin
3:20 PM Panel Discussion: Stacy Lindborg; Chito Hernandez; Ivan
Chan; Lei Wang; Naiitee Ting
4:30 PM Floor Discussion.

Session 71: General Developments in Nonclinical Statistics
(Top Contributed)
Room: CONFERENCE ROOM 4Q, FOURTH FLOOR
Organizer: Jorge Quirzo and Yanbing Zheng; Merck and Abbvie.
Chair: Yanbing Zheng; Abbvie Inc.
2:50 PM Bayesian hierarchical model estimation and comparison of
immunogenicity assay cut-points from pre-study
• Dave LeBlond¹, Rong Zeng², Lu Xu² and Robert Singer³.
¹Statistical Consultant ²OncoMed Pharmaceuticals ³Robert Singer Consulting
3:15 PM Use of High Confidence Sampling Plans for Product Assessment
Mark Johnson. Abbvie Inc.
3:40 PM Exact Trend Test on Comparing Tumor Incidence in Transgenic Mouse Carcinogenicity Studies
Lei Shu. Astellas Pharma
4:05 PM A machine learning approach to biomarker discovery from
high dimensional omics data
Jie Cheng. Abbvie Inc.
Tuesday, June 27. 10:30 AM - 12:10 PM

Session 74: Statistical and Computational Challenges for Single-cell Sequencing (Invited)
Room: CONFERENCE ROOM 4J, FOURTH FLOOR
Organizer: An, Lingling; University of Arizona.
Chair: An, Lingling; University of Arizona.

10:30 AM SClnorm: A quantile-regression based approach for robust normalization of single-cell RNA-seq data
*Rhonda Bacher1, Li-Fang Chu2, Audrey Gasch1, James Thomson3, Ron Stewart2, Michael Newton1 and Christina Kendziorski1.
1University of Wisconsin-Madison
2Morgridge Institute for Research

10:55 AM Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning
*Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti and Serafin Batzoglou.
Stanford University

11:20 AM Global Prediction of Chromatin Accessibility Using RNA-seq from Single Cell and Small Number of Cell
Weiqiang Zhou, Zhicheng Ji and *Hongkai Ji.
Department of Biostatistics, Johns Hopkins Univ.

11:45 AM Expression Recovery in Single Cell RNA Sequencing
Mo Huang, Mingyao Li and *Nancy Zhang.
University of Pennsylvania

12:10 PM Floor Discussion.

Session 75: Several advanced topics in statistical metagenomics (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Chen, Jun; Mayo Clinic.
Chair: Chen, Jun; Mayo Clinic.

10:30 AM Regularized Hotelling’s Test for High Dimensional Paired Microbiome Compositional Data
*Ni Zhao1, Xiang Zhan2 and Michael Wu2.
1Johns Hopkins University
2Fred Hutchinson Cancer Research Center

10:55 AM A kernel independence test for microbiome community-level association analysis
*Xiang Zhan1, Anna Plantinga2, Ni Zhao3 and Michael Wu1.
1Fred Hutch Cancer Research Center
2University of Washington
3John Hopkins University

11:20 AM Trace Evidence Analysis Utilizing Metagenome Sequence Data
*Kyle Carter, Meng Lu and Lingling An.
University of Arizona

11:45 AM A Marginalized Two-Part Beta Regression Model for Microbiome Compositional Data
*Haitao Chai1, Hongmei Jiang2, Lu Lin1 and Lei Liu2.
1Shandong University
2Northwestern University

12:10 PM Floor Discussion.
Session 76: Recent Advancements in Personalized Medicine (Invited)
Room: CONFERENCE ROOM 4I, FOURTH FLOOR
Organizer: Chiang, Alan; Eli Lilly and Company.
Chair: Chiang, Alan; Eli Lilly and Company.

10:30 AM Empirical Likelihood for Comparing Two Survival Functions
Hsin-wen Chang. Academia Sinica

10:55 AM Subgroup identification with latent Dirichlet allocation
Hyunbo Chan. Purdue University

11:20 AM Adjusting treatment effect estimated from an exploratory biomarker analysis using resampling methods
Aimee Wang, Tianle Hu and Honglu Liu. Eli Lilly and Company

11:45 AM Floor Discussion.

Session 77: Statistical Tests for Data Measured on Taxonomic Trees in Microbiome Studies (Invited)
Room: CONFERENCE ROOM 4H, FOURTH FLOOR
Organizer: Li, Hongzhe; University of Pennsylvania.
Chair: Michael Sohn; University of Pennsylvania.

10:30 AM A General Framework for Association Analysis of Microbial Communities on a Taxonomic Tree
Zheng-Zheng Tang1, Guanhua Chen1, Alexander Aleskseyenko2 and Hongze Li3. 1Vanderbilt University School of Medicine 2Medical University of South Carolina 3University of Pennsylvania School of Medicine

10:55 AM The Phylogenetic LASSO and the Microbiome
Stephen T Rash1, Jessmyn Niergarth1 and Peter T Kim. 1University of Guelph

11:20 AM Kernel-penalized regression models for taxon-specific associations in microbiome studies
Timothy Randolph1, Sen Zhao2 and Ali Shojaie3. 1Fred Hutchinson Cancer Research Center 2University of Washington

11:45 AM Data exploratory methods for microbiome data analysis
Hong Gu and Toby Kenney. Dalhousie University

12:10 PM Floor Discussion.

Session 78: Complex and Large Data Analysis (Invited)
Room: WILLIFORD A, THIRD FLOOR
Organizer: Li, Lexin; UC Berkeley.
Chair: Shou, Haochang; University of Pennsylvania.

10:30 AM Jointly analysis of the NMR and MS data by simultaneous functional data deconvolution
Yiwen Liu, Wenzuan Zhong, Ping Ma and Arthur Edison. University of Georgia

10:55 AM Functional Joint Model with application of Alzheimer’s disease
Kan Li and Sheng Luo. The University of Texas at Houston

11:20 AM Bayesian Functional Graphical Models
Lin Zhang1, Veera Baladandayuthapani2, Francesco Verse3 and Jeffrey Morris4. 1University of Minnesota 2University of Texas MD Anderson Cancer Center 3University of Oklahoma Health Sciences Center 4University of Texas MD Anderson Cancer Center

11:45 AM Floor Discussion.

Session 79: Inference for Big and Complex Data (Invited)
Room: CONFERENCE ROOM 4P, FOURTH FLOOR
Organizer: Lin, Wei; Peking University.
Chair: Xiaojun Mao; Iowa State University.

10:30 AM Differential Network Analysis via Lasso Penalized D-Trace Loss
Huili Yuan, Ruibin Xi, Chong Chen and Minghua Deng. Peking University

10:55 AM Scalable Interpretable Multi-Response Regression via SEED
Zemin Zheng. University of Science and Technology of China

11:20 AM gCoda: conditional dependence network inference for compositional data
Huaying Fang1, Chengcheng Huang2, Hongyu Zhao3 and Minghua Deng4. 1Stanford University School of Medicine 2China Meteorological Administration 3Yale School of Public Health 4Peking University

11:45 AM Identifiability and Inference of Causal Effects with High-Dimensional and Invalid Instruments
Changjing Wu, Minghua Deng and Wei Lin. Peking University

12:10 PM Floor Discussion.

Session 80: Quantile regression and inference for high-dimensional problems (Invited)
Room: CONFERENCE ROOM 4R, FOURTH FLOOR
Organizer: Peng, Limin; Emory University.
Chair: Peng, Limin; Emory University.

10:30 AM Principal Quantile Regression for Sufficient Dimension Reduction with Heteroscedasticity
Chong Wang1, Seung Jun Shin2 and Yichao Wu3. 1North Carolina State University 2Korea University

10:55 AM Variable Selection for High-Dimensional Additive Quantile Regression
Ben Sherwood1 and Adam Maidman2. 1University of Kansas 2University of Minnesota

11:20 AM Censored quantile regression in high dimensional survival data
Qi Zheng1, Limin Peng2 and Xuming He3. 1University of Louisville 2Emory University 3University of Michigan

11:45 AM Floor Discussion.

Session 81: Recent Advances on Statistical Modeling for Cancer Genomics Data (Invited)
Room: CONFERENCE ROOM 5G, FIFTH FLOOR
Organizer: Qin, Li-Xuan; Memorial Sloan Kettering Cancer Center.
Chair: Liu, Jin; Duke-NUS Singapore.
Session 82: Statistical considerations in Multi-regional trials (Invited)
Room: WILLIFORD B, THIRD FLOOR
Organizer: Tsong, Yi; FDA.
Chair: Yi Tsong; FDA.
10:30 AM Some Recent Advances on Statistical Approaches for Planning Multi-Regional Clinical Trials
H.M. James Hung, Xiangmin Zhang and Sue-Jane Wang.
Food and Drug Administration
10:55 AM Multi-regional Biosimilarity Studies
Yu-Wei Chang1 and Xia Qi2. 1Boehringer Ingelheim
Temple University
11:20 AM Design and Evaluation of Multiregional Clinical Trials with Heterogeneous Variability across Regions
Chin-Fu Hsiao and Chieh Chiang. National Health Research Institutes
11:45 AM Floor Discussion.

Session 83: Combining Endpoints in Clinical Trials (Invited)
Room: CONFERENCE ROOM 5H, FIFTH FLOOR
Organizers: Wang, Chenkun/DeSouza, Cynthia; Vertex Pharmaceuticals.
Chair: Chenkun Wang; Vertex Pharmaceuticals.
10:30 AM Weighted win loss approach for analyzing prioritized outcomes
Xiaodong Luo1, Junshan Qiu2, Steven Bai2 and Hong Tian3. 1Sanofi US 2US Food and Drug Administration 3Janssen Research and Development
10:55 AM Generalized pairwise comparisons for prioritized outcome measures
Marc Buyse1, Jullien Peron2, Joris Giai3 and Pascal Roy3. 1IDDI Inc., San Francisco 2Medical Oncology Department, CH Lyon Sud, Lyon 3Hospices Civils de Lyon, Service de Biostatistique
11:20 AM Measures of Clinical Benefit in Immuno-Oncology Studies
Laping Zhao, Yifan Huang, Yiduo Zhang, Kelvin Shi, Wenwei Huang and Pralay Mukhopadhyay. AstraZeneca
11:45 AM Analyzing Multiple Endpoints Simultaneously in Randomized Clinical Trials using GPC
Joris Giai1, Julien Peron1, Delphine Maucort-Boulch1, Pascal Roy3 and Marc Buyse2. 1Hospices Civils de Lyon / CNRS UMR 5558 / UCBL 2International Drug Development Institute (IDDI)
12:10 PM Floor Discussion.

Session 84: Biomarker development: Ideas and practicalities
(Invited)
Room: CONFERENCE ROOM 5F, FIFTH FLOOR
Organizer: Wang, Hao; Johns Hopkins University.
Chair: Wang, Hao; Johns Hopkins University.
10:30 AM Biomarker Development: Ideals and Practicalities
Karla Ballman. Weil Cornell Medicine
10:55 AM Designing the trial of anti-PD-1 therapy in tumors with mismatch repair deficiency: Innovative study
Hao Wang. Johns Hopkins University School of Medicine
11:20 AM Biomarker development: Ideas and practicalities
Bruce Trock. Johns Hopkins University
11:45 PM Floor Discussion.

Session 85: Statistical Analysis for Non-normal Data (Invited)
Room: WILLIFORD C, THIRD FLOOR
Organizer: Wang, Jianming; Celgene.
Chair: Carl Di Casoli; Celgene.
10:30 AM High Dimensional Variable Selection for Censored Quantile Regression
Naveen Naidu Narisetty1 and Xuming He2. 1University of Illinois at Urbana Champaign 2University of Michigan
10:55 AM Measures of Income Inequality Focusing on the Lower and Middle Classes
11:20 AM Under-dispersion models: models that are "under the radar"
Kimberly Sellers1 and Darcy Morris2. 1Georgetown University, and U.S. Census Bureau 2U.S. Census Bureau
11:45 AM A class of semiparametric transformation models and their applications
Guoqing Diao. George Mason University
12:10 PM Floor Discussion.

Session 86: Statistical methods in oncology trials (Invited)
Room: ASTORIA ROOM, THIRD FLOOR
Organizer: Wang, Jianming; Celgene.
Chair: Carl Di Casoli; Celgene.
10:30 AM Predicting Data Cut-off Date in Event-Driven Clinical Trials When There is a Lag in Adjudication
Jianming Wang and Lei Fu. Celgene Corporation
10:55 AM Correcting Treatment Effect for Treatment Switching in Randomized Oncology Trials with a Generalize
11:20 AM The Utility of the Yang-Prentice Model when the Cox Pro-
portional Hazards Assumption is Violated
Carl Di Casoli. Celgene Corporation

11:45 AM Floor Discussion.

Session 87: Mixed-effects models in genomics, health and
psychology (Invited)
Room: CONFERENCE ROOM 5I, FIFTH FLOOR
Organizer: Wang, Jiebiao; University of Chicago.
Chair: Qin, Xu; University of Chicago.

10:30 AM A unified powerful set-based test for sequencing data analy-
sis of GxE interactions
♦ Yu-Ru Su, Chong-Zhi Di and Li Hsu. FRED HUTCHINSON CANCER RESEARCH CENTER

10:55 AM Modeling of Between- and Within-Subject Variances Using
Mixed Effects Location Scale Models
♦ Donald Hedeker1 and Robin Mermelstein2. 1University of Chicago 2University of Illinois at Chicago

11:20 AM On estimating within-person relations using mixed-effects modeling
♦ Lijuan Wang1, Qian Zhang2, Scott Maxwell1 and Cindy Bergeman1. 1University of Notre Dame 2Florida State University

11:45 AM Modeling non-normal distributions in mixed-effects and multilevel models
Zhiyong Zhang. University of Notre Dame

12:10 PM Floor Discussion.

Session 88: Making Sense of Big Omics Data "C Recent Advances in Statistical and Computational Methods (Invited)
Room: CONFERENCE ROOM 5E, FIFTH FLOOR
Organizer: Ding, Ying; University of Pittsburgh.
Chair: Ding, Ying; University of Pittsburgh.

10:30 AM Statistical methods for Hi-C data to study DNA structure
♦ Yan Li1, Ming Hu2, Joshua Martin1, Zheng Xu3, Bing Ren4 and Karen Mohlke5. 1University of North Carolina, Chapel Hill 2Cleveland Clinic Foundation, 3University of Nebraska, Lincoln, 4University of California, San Diego

10:55 AM A Dirichlet mixture model for clustering droplet-based single cell transcriptomic data
Zhe Sun1, Ting Wang1, Ming Hu2 and ♦ Wei Chen1. 1University of Pittsburgh 2Cleveland Clinic Foundation

11:20 AM Bayesian hierarchical model with multi-layer overlapping group structure in omics applications
Li Zhu and ♦ George Tseng. Department of Biostatistics, Univ. of Pittsburgh

11:45 AM A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome
Anthony Schmitt1, ♦ Ming Hu2 and Bing Ren1. 1University of California San Diego 2Cleveland Clinic

12:10 PM Floor Discussion.

Session 89: Statistical Applications in Population Health Sciences (Invited)
Room: CONFERENCE ROOM 5A, FIFTH FLOOR
Organizer: Hsiao, Chin-Fu; National Health Research Institute Taiwan.
Chair: Toshimitsu Hamasaki; National Cerebral and Cardiovascular Center Japan.

10:30 AM Using genetic markers in risk prediction and pharmacoge-
nomics studies for lung cancer
Chao Hsiung. National Health Research Institutes

10:55 AM Longitudinal studies of children and adolescent behaviors
Hsing-Ti Chang. National Health Research Institutes

11:20 AM Pharmacovigilance studies of psychotrophic agents’ opportu-
nities and pitfalls of big data
♦ Hui-Ju Tsai1, Chi-Shin Wu2, Yu-Ting Tsai1 and Ya-Wen Huang3. 1National Health Research Institutes 2National Taiwan University Hospital

11:45 AM Methods and Theory for Spatial Change Set Analysis
Pei-Sheng Lin. National Health Research Institutes

12:10 PM Floor Discussion.

Session 90: New Advances on Nonparametric Methods for
High-dimensional Data (Top Contributed)
Room: CONFERENCE ROOM 4Q, FOURTH FLOOR
Organizer: Cao, Guanqun; Auburn University.
Chair: Li, Xinyi; Iowa State University.

10:30 AM Tree-based model for longitudinal data
♦ Peng Wang and Brittany Green. University of Cincinnati

10:50 AM Partially Linear Single-Index Models in Ultra-high Dimension
♦ Shan Yu, Lily Wang and Dan Nettleton. Iowa State University

11:10 AM Robust high-dimensional regression and variable selection
♦ Yichen Qin1, Shaobo Li1, Yang Li2 and Yan Yu1. 1University of Cincinnati 2Renmin University of China

11:30 AM Fast Covariance Estimation for Sparse Functional Data
Luo Xiao1, ♦ Cai Li1, William Checkley 2 and Ciprian Crainiceanu2. 1North Carolina State University 2Johns Hopkins University

11:50 AM Modeling Microbial Abundance by Zero-Inflated Semi-parametric Gamma Mixed-Effect Models
♦ Hongmei Yang, David Topham Xing Qiu, Mary Caserta and Gloria Pyhuber. URM

12:10 PM Floor Discussion.

Session 91: Student session (II) (Student Winner Session)
Room: CONFERENCE ROOM 5B, FIFTH FLOOR
Chair: Haoran Li; University of California Davis.

10:30 AM AN ADAPTABLE GENERALIZATION OF HOTELLING’S T2 TEST IN HIGH DIMENSION
♦ HAORAN Li1, Alexander Aue1, Debasis Paul1, Jie Peng1 and Pei Wang2. 1University of California Davis 2School of Medicine at Mount Sinai

12:10 PM Floor Discussion.
Scientific Program (Presenting Author) Tuesday, June 27. 1:30 PM - 3:10 PM

Session 92: New Advances in Modeling of Functional Data (Invited)
Room: CONFERENCE ROOM 4P, FOURTH FLOOR
Organizer: Ogden, R.Todd; Columbia University.
Chair: Hongxia Zhu; Virginia Tech.

1:30 PM High-Dimensional Function-on-Scalar Regression
Matthew Reimherr1 and Alice Parodi2.
1Penn State University
2Politecnico di Milano

1:55 PM Image-on-Scalar Regression via Bivariate Penalized Splines
Guannan Wang1 and Lily Wang2.
1College of William & Mary
2Iowa State University

2:20 PM A Multi-Dimensional Functional Principal Components Analysis of EEG Data
Kyle Hasenstab, Aaron Scheffler, Donataello Telesca,
Catherine Sugar, Shafali Jeste, Charlotte DiStefano and
Damla Senturk.
UCLA

2:45 PM Floor Discussion.

Session 93: Statistical and Computational Approaches for High-throughput Sequencing Data Analysis (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Shen, Ronglai; Memorial Sloan Kettering Cancer Center.
Chair: Shen, Ronglai; Memorial Sloan Kettering Cancer Center.

1:30 PM Utilizing Patient-Level Characteristics for Identification of Cancer Driver Genes
Ho-Hsiang Wu1, Nilanjan Chatterjee2 and Bin Zhu1.
1National Cancer Institute
2Johns Hopkins University

1:55 PM Reproducible RNA-seq analysis with recount2
Leonardo Collado Torres1, Abhinav Nelllore2, Kai, Shannon
Kammers, Ellis2, Margaret, Kasper Taub, Hansen3, Andrew
Jaffe1, Ben Langmead4 and Jeff Leek5.
1Lieber Institute for Brain Development
2Comp. Bio. Program and Dept. of Surgery, OHSU
3Department of Biostatistics, JHSPH
4Department of Computer Science, JHU

2:20 PM TSNet: a new method for constructing tumor specific gene
co-expression networks
Francesca Petralia, Li Wang and Pei Wang.
Icahn School of Medicine at Mt Sinai, New York

2:45 PM Floor Discussion.

Session 94: Non-clinical and biosimilar (Invited)
Room: WILLIFORD B, THIRD FLOOR
Organizer: Tsong, Yi; FDA.
Chair: Meiyu Shen; FDA.

1:30 PM Statistical methods for equivalence assessment of analytical biosimilar
Yi Tsong, Meiyu Shen, Yu-Ting Weng and Chao Wang.
CDER, FDA

1:55 PM Methods and Applications of Percentile Estimation
Qi Xia1, Yi Tsong2 and Yu-Ting Weng3.
1Temple University
2FDA

2:20 PM Statistical Considerations in Demonstrating CMC Analytical Similarity for a Biosimilar Product
Richard Burdick1, Neal Thomas2 and Aili Cheng3.
1Elion Labs, Arizona State University
2Pfizer, Statistical Research and Consulting Center
3Pfizer, Pharmaceutical Sciences and Manufacturing

2:45 PM SAMPLE SIZE REQUIREMENT FOR ANALYTICAL BIOSIMILARITY ASSESSMENT
Tianhua Wang, Yi Tsong and Meiyu Shen.
FDA

3:10 PM Floor Discussion.

Session 95: Statistical Methods for Modelling Data Complexity in Genomics Studies (Invited)
Room: CONFERENCE ROOM 4H, FOURTH FLOOR
Organizer: Tzeng, Jung-Ying; NC State University.
Chair: Lin Chen; NC State University.

1:30 PM Fast Bayesian Variable Screenings for Binary Response Regressions with Small Sample Size
Sheng-Mao Chang1, Jung-Ying Tzeng2 and Ray-Bing
Chen2.
1National Cheng Kung University
2North Carolina State University

1:55 PM Multivariate selection models for labelling-based proteomics data with non-ignorable missingness
Lin Chen1, Jiebiao Wang1, Pei Wang2 and Don Hedeker3.
1University of Chicago
2Mt Sinai Hospital

2:20 PM Joint Modeling and Analysis of Microbiome with Other Omics Data
Xiang Zhan1, Anna Plantinga2, Ni Zhao3 and Michael
Wu4.
1Fred Hutchinson Cancer Research Center
2University of Washington
3The Johns Hopkins University

2:45 PM A Non-parametric Framework for High Dimensional Multi-
Modal Data With Applications to Imaging Genetics
Zhaoxia Yu1, Dustin Pluta1, Tong Shen1 and Hernando
Ombao2.
1University of California, Irvine
2University of California, Irvine; KAUST

3:10 PM Floor Discussion.

Session 96: Recent advances in early phase oncology trials using drug combinations (Invited)
Room: CONFERENCE ROOM 4R, FOURTH FLOOR
Organizer: Wang, Ling; Takeda.
Chair: Wang, Ling; Takeda.

1:30 PM Bayesian dose-finding designs for drug combinations based on
time to response and toxicity outcomes
Xiao Su1 and Yisheng Li2.
1The University of Texas School of Public Health
2The University of Texas MD Anderson Cancer Center
1:55 PM A curve-free Bayesian decision-theoretic design for two-agent phase I trial and its extension
Yong Lu1, Bee Leng Lee2 and Hua Jin1. 1Stanford University 2San Jose State University, 3South China Normal University
2:20 PM Discussant: Yuan Ji
2:45 PM Floor Discussion.

Session 97: Advanced Statistical Methods for Complicated High Dimensional Data (Invited)
Room: CONFERENCE ROOM 5G, FIFTH FLOOR
Organizer: Wang, Pei; Icahn School of Medicine at Mount Sinai.
Chair: I-Ping Tu; The Institute of Statistical Science, Academia Sinica.
1:30 PM A new approach for identifying and removing the cell-cycle effect from single-cell RNA-seq data
Martin Barron and Jun Li. University of Notre Dame
1:55 PM Robust Normalization of High-throughput Proteomics Experiments
Hua Tang, Meng Wang, Huaying Fang, Lihua Jiang, Shin Lin and Michael Snyder. Stanford University
2:20 PM On Consistency of Graph-based Semi-supervised Learning
Chengan Du and Yanpeng Zhao. George Mason University
2:45 PM QRank: A novel quantile regression tool for eQTL discovery
Xiaoyu Song, Gen Li, Zhenwei Zhou, Xianling Wang, Jiuliana Ionita-Laza and Ying Wei. Columbia University
3:10 PM Floor Discussion.

Session 98: Advances in Low-rank Modeling and Its Estimation (Invited)
Room: CONFERENCE ROOM 5H, FIFTH FLOOR
Organizer: Wong, Raymond K.W.; Iowa State University.
Chair: Zhang, Xiaoke; University of Delaware.
1:30 PM Recent Advances in Collaborative Ranking
Cho-Jui Hsieh. UC Davis
1:55 PM Nonparametric Operator-Regularized Covariance Function Estimation
Raymond K. W. Wong1 and Xiaoke Zhang2. 1Iowa State University 2University of Delaware
2:20 PM Parsimonious matrix-variate regressions through envelope models
Shanshan Ding1 and Dennis Cook2. 1University of Delaware 2University of Minnesota
2:45 PM MM Algorithms For Variance Components Models
Hua Zhou1, Liyvi Hu2, Jin Zhou1 and Kenneth Lange1. 1UCLA 2NCSU 3University of Arizona
3:10 PM Floor Discussion.

Session 99: Recent Advancement and Applications in Multiplicity Adjustments (Invited)
Room: CONFERENCE ROOM 5F, FIFTH FLOOR
Organizer: Xu, Tu; AbbVie Inc.
Chair: Xu, Tu; AbbVie Inc.
1:30 PM Large-Scale Heterogeneity Testing under Sparsity
Xiang Lu and Guang Cheng. Department of Statistics, Purdue University
1:55 PM A Unified Framework For Weighted Parametric Multiple Test Procedures
Dong Xi1, Ekkehard Glimm2, Willi Maurer2 and Frank Bretz2. 1Novartis Pharmaceuticals Corporation 2Novartis AG
2:20 PM Robustness of Multiple Testing Procedures in Confirmatory Clinical Trials
Michael Rosenblum, Chenguang Wang and Yuchen Yang. Johns Hopkins University
2:45 PM Floor Discussion.

Session 100: Design and analysis of cancer immunotherapy trials (Invited)
Room: ASTORIA ROOM, THIRD FLOOR
Organizer: Xu, Zhenzhen; FDA.
Chair: Xu, Zhenzhen; FDA.
1:30 PM Practical considerations in designing immunotherapy trials with delayed treatment effects
Alan Chiang. Eli Lilly and Company
1:55 PM Achieving Optimal Power of Log-rank Test with Random Treatment Time-lag Effect
Zhenzhen Xu1, Yongsoek Park2, Boguang Zhen1 and Bin Zhu3. 1FDA/CBER 2University of Pittsburgh 3National Cancer Institute
2:20 PM Flexible Survival Model and Sample Size Determination in Cancer Vaccine Studies
Takahiro Hasegawa. Shionogi & Co., Ltd.
2:45 PM Discussant: Boguang Zhen
3:10 PM Floor Discussion.

Session 101: Design and practical considerations for dose selection trials (Invited)
Room: CONFERENCE ROOM 5I, FIFTH FLOOR
Organizer: Zhang, Yanqiong; Agios Pharmaceuticals.
Chair: Zhang, Yanqiong; Agios Pharmaceuticals.
1:30 PM Dose Finding Based on Integrated and Adaptive Approaches
Lei Nie and Rajeshwari Sridhara. The US FDA
1:55 PM Practical Experiences with Phase 1 Dose Escalation Studies
Inna Perevozskaya and Yuehui Wu. GSK
2:20 PM Application of Bayesian Methods in Oncology Dose Escalation Studies with Late Onset Toxicity
Li Liu. Sanofi
2:45 PM Longitudinal Data Generation Based on Bootstrapped Samples from an Observational Study with Application
Greg Cicconetti1, Deli Wang1 and Weining Robieson. 1Abbvie
3:10 PM Floor Discussion.
Session 102: Novel Statistical Methods for Analysis of Complex Biological Data (Invited)
Room: CONFERENCE ROOM 5E, FIFTH FLOOR
Organizer: Zhao, Dave; UIUC.
Chair: Zhang, Yu; Penn State University.
1:30 PM Accounting for Within-gene Correlation Structure in RNA-seq Differential Expression Analysis
Yet Nguyen and \*Dan Nettleton. Iowa State University
1:55 PM Tensor-on-tensor Regression
Eric Lock. University of Minnesota
2:20 PM Bayesian Semiparametric Mixed Effects Markov Chains
\*Abhra Sarkar and David Dunson. Duke University
2:45 PM Computation of Ancestry Scores with Mixed Families and Unrelated Individuals
\*Yi-Hui Zhou\*1, J.S. Marron\*2 and Fred Wright\*1. \*1North Carolina State University \*2University of North Carolina
3:10 PM Floor Discussion.

Session 103: TBD (Invited)
Room: WILLIFORD A, THIRD FLOOR
Organizer: Zhao, Zhigen; Temple University.
Chair: Zhao, Zhigen; Temple University.
1:30 PM Empirical Bayes and a flexible multiple-testing framework for large-scale simultaneous inference
Zhi Wei. New Jersey Institute of Technology
1:55 PM Flexible Spectral Methods for Community Detection in Networks
Pengsheng Ji. University of Georgia
2:20 PM A Scalable Empirical Bayes Approach to Variable Selection in Generalized Linear Models
\*Haim Bar\*1, James Booth\*2, Martin Wells\*2 and Kangyan Liu\*1. \*1University of Connecticut \*2Cornell University
2:45 PM Floor Discussion.

Session 104: Analysis of High-Dimensional Data and Applications (Invited)
Room: CONFERENCE ROOM 5A, FIFTH FLOOR
Organizer: Yuan, Ming; University of Wisconsin.
Chair: Yuan, Ming; University of Wisconsin.
1:30 PM Intrinsic Entropies of Log-concave Distributions
Varun Jog. University of Wisconsin - Madison
1:55 PM High Dimensional Minimum Variance Portfolio under Factor Model
Yi Ding, \*Yingying Li and Xinghua Zheng. Hong Kong University of Science and Technology
2:20 PM Testing and Scoring High-dimensional Covariance Matrices When Heteroscedasticity is Present
Xinxin Yang\*1, \*Xinghua Zheng\*3, Jiaqi Chen\*2 and Hua Li\*3. \*1HKUST \*2Harbin Institute of Technology \*3Chang Chun University
2:45 PM Canonical Correlation Analysis: New Losses And New Rates
Zhuang Ma\*1 and \*Xiaodong Li\*2. \*1University of Pennsylvania \*2University of California, Davis
3:10 PM Floor Discussion.

Session 105: Recent advances in clinical trial design and analysis with time to event endpoints (Invited)
Room: CONFERENCE ROOM 4Q, FOURTH FLOOR
Organizer: Yang, Fang and Zhu, Hong; Novartis.
Chair: Zhu, Hong; UT Southwestern Medical Center.
1:30 PM The Win-ratio Statistic in Clinical Trials with Multiple Types of Event
David Oakes. University of Rochester
1:55 PM Semi-parametric Density Ratio Modeling of Survival Data From a Prevalent Cohort
\*Hong Zhu\*1, Jing Ning\*2. Yu Shen\*2 and Jing Qin\*3. \*1University of Texas Southwestern Medical Center \*2University of Texas MD Anderson Cancer Center \*3National Institute of Allergy & Infectious Disease
2:20 PM A Bayesian Nonparametric Model to Predict the Time of Hospital Readmission
Song Zhang. University of Texas Southwestern Medical Center
2:45 PM Floor Discussion.

Session 106: Methods and applications in clustered and high dimensional data (Invited)
Room: CONFERENCE ROOM 4J, FOURTH FLOOR
Organizer: Mei-Ling Ting Lee; University of Maryland.
Chair: Wei-Liang Qiu; Harvard University.
1:30 PM Bayesian multivariate skew meta-regression models for individual patient data
Joseph Ibrahim\*1, Sungduk Kim\*2, \*Ming-Hui Chen\*3, Arvind Shah\*4 and Jianxin Lin\*5. \*1University of North Carolina \*2National Cancer Institute \*3University of Connecticut \*4Merck Research Laboratories
1:55 PM Dimension Reduction in High Dimensional Multivariate Time Series Analysis
Zeda Li and \*William W.S. Wei. Dept. of Statistical Science, Temple University
2:20 PM Semiparametric Monitoring test based on clustered data
\*Jiahua Chen\*3, Pengfei Li\*2 and Yukun Liu\*3. \*1Dept of Statistics, University of British Columbia \*2Dept of Statistics, University of Waterloo \*3School of Mathematics and Statistics, ECNU
2:45 PM Automatic Region-wise Spatially Varying Coefficient Regression Model
Shuo Chen and \*Chengsheng Jiang. University of Maryland, College Park
3:10 PM Floor Discussion.
1:30 PM A Calibrated Power Prior Approach to Borrow Information from Historical Data with Application to Bio
*Haitao Pan¹, Ying Yuan¹ and Jielai Xia². ¹The University of Texas MD Anderson Cancer Center ²Fourth Military Medical University

1:55 PM A Nonparametric Bayesian Basket Trial Design
*Yaxuan Xu², Peter Mueller², Apostolia Tsimeridou¹ and Donald Berry³. ¹Johns Hopkins University ²The University of Texas at Austin ³The University of Texas MD Anderson Cancer Center

2:20 PM Group Sequential Design Comparing Multiple Treatments to a Common Control
*Jon Steingrimsson, Joshua Betz and Michael Rosenblum. Johns Hopkins University

2:45 PM Two Stage Drop the Loser Design in Clinical Trials with Response Adaptive Randomization
*Hongjian Zhu¹, Jin Piao², Jack Lee² and Feifang Hu³. ¹University of Texas Health Science Center at Houston ²University of Texas MD Anderson Cancer Center ³George Washington University

3:10 PM Floor Discussion.

Session 108: Topics on Applied Statistics (I) (Contributed)
Room: CONFERENCE ROOM 4I, FOURTH FLOOR
Chair: Haiming Zhou; Northern Illinois University.

1:30 PM Fitting Bayesian Spatial Survival Models Using R
*Haiming Zhou¹ and Timothy Hanson². ¹Northern Illinois University ²University of South Carolina

1:45 PM A Robust Two-stage Design Identifying the Optimal Biological Dose for Phase I/II Clinical Trials
*Yong Zang¹ and Jack Lee². ¹Indiana University ²MD Anderson Cancer Center

2:00 PM A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies
*Jingjing Yang¹, Lara Fritsche¹, Xiang Zhou¹ and Goncalo Abecasis¹. ¹University of Michigan ²Norwegian University of Science and Technology

2:15 PM A Simple Method for Bayesian Variable Selection Based on Parameter Estimates, with Application to Me
*Grace Yoon, Wenxin Jiang and Lei Liu. Northwestern University

2:30 PM Statistical Considerations in Using Meta-analysis for Regulatory Decision Making for Medical Devices
Qin Li. FDA Center for Devices and Radiological Health

2:45 PM Parallelized Adaptive Rejection Metropolis Sampling on the Utility of Graphics Cards
Yuefeng Wu. University of Missouri St. Louis

3:00 PM Floor Discussion.

Session 109: Topics on structured models (Contributed)
Room: CONFERENCE ROOM 4I, FOURTH FLOOR
Chair: Lifeng Lin; University of Minnesota.

1:30 PM On Evidence Cycles in Network Meta-analysis
*Lifeng Lin, Haitao Chu and James Hodges. University of Minnesota

1:50 PM Dynamic Transelliptical Graphical Models for Sparse Precision Matrix Estimation
*Tzu-Chun Wu and Emily Kang. University of Cincinnati

2:10 PM Clinical Response Prediction with Logistical Model Adjusted by Tree Regression
Dion Chen. Janssen R&D, LLC

2:30 PM Characterizing Spatial Dependence on Stream Networks — Bayesian Hierarchical Model Approximation
*Yingying Liu and Kate Cowles. The University of Iowa

2:50 PM Graphical Horseshoe Model for Inverse Covariance Estimation
*Yunfan Li, Anindya Bhadra and Bruce Craig. Purdue University

3:10 PM Floor Discussion.

Session 110: Topics on structured models (Contributed)
Room: CONFERENCE ROOM 4I, FOURTH FLOOR
Chair: Chen, Hao; University of California Davis.

1:30 PM On Evidence Cycles in Network Meta-analysis
*Lifeng Lin, Haitao Chu and James Hodges. University of Minnesota

1:50 PM Dynamic Transelliptical Graphical Models for Sparse Precision Matrix Estimation
*Tzu-Chun Wu and Emily Kang. University of Cincinnati

2:10 PM Clinical Response Prediction with Logistical Model Adjusted by Tree Regression
Dion Chen. Janssen R&D, LLC

2:30 PM Characterizing Spatial Dependence on Stream Networks — Bayesian Hierarchical Model Approximation
*Yingying Liu and Kate Cowles. The University of Iowa

2:50 PM Graphical Horseshoe Model for Inverse Covariance Estimation
*Yunfan Li, Anindya Bhadra and Bruce Craig. Purdue University

3:10 PM Floor Discussion.

Session 111: The challenges of non-constant hazard ratio: delayed treatment effect, treatment dilution and treatment crossover (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Luo, Xiaodong; Sanofi.
Chair: Quan, Hui; Sanofi.

3:30 PM Change Point Analysis of Correlation in Non-stationary Time Series
Holger Dette¹, Weichi Wu² and Zhou Zhou³. ¹Ruhr-University Bochum ²University College London ³University of Toronto

3:55 PM Multi-Sequence Segmentation Using Score Tests
Hock Peng Chan. National University of Singapore

4:20 PM Computationally Efficient Methods for Multivariate Change-point Analysis with Industrial Application
Idris Eckley. Lancaster University

4:45 PM DNA Copy Number Profiling Using Single-cell Sequencing
Xuefeng Wang. H. Lee Moffitt Cancer Center & Research Institute

5:10 PM Floor Discussion.

Session 111: The challenges of non-constant hazard ratio: delayed treatment effect, treatment dilution and treatment crossover (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Luo, Xiaodong; Sanofi.
Chair: Quan, Hui; Sanofi.

3:30 PM Change Point Analysis of Correlation in Non-stationary Time Series
Holger Dette¹, Weichi Wu² and Zhou Zhou³. ¹Ruhr-University Bochum ²University College London ³University of Toronto

3:55 PM Multi-Sequence Segmentation Using Score Tests
Hock Peng Chan. National University of Singapore

4:20 PM Computationally Efficient Methods for Multivariate Change-point Analysis with Industrial Application
Idris Eckley. Lancaster University

4:45 PM DNA Copy Number Profiling Using Single-cell Sequencing
Xuefeng Wang. H. Lee Moffitt Cancer Center & Research Institute

5:10 PM Floor Discussion.

Session 111: The challenges of non-constant hazard ratio: delayed treatment effect, treatment dilution and treatment crossover (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Luo, Xiaodong; Sanofi.
Chair: Quan, Hui; Sanofi.

3:30 PM Change Point Analysis of Correlation in Non-stationary Time Series
Holger Dette¹, Weichi Wu² and Zhou Zhou³. ¹Ruhr-University Bochum ²University College London ³University of Toronto

3:55 PM Multi-Sequence Segmentation Using Score Tests
Hock Peng Chan. National University of Singapore

4:20 PM Computationally Efficient Methods for Multivariate Change-point Analysis with Industrial Application
Idris Eckley. Lancaster University

4:45 PM DNA Copy Number Profiling Using Single-cell Sequencing
Xuefeng Wang. H. Lee Moffitt Cancer Center & Research Institute

5:10 PM Floor Discussion.

Session 111: The challenges of non-constant hazard ratio: delayed treatment effect, treatment dilution and treatment crossover (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Luo, Xiaodong; Sanofi.
Chair: Quan, Hui; Sanofi.

3:30 PM Change Point Analysis of Correlation in Non-stationary Time Series
Holger Dette¹, Weichi Wu² and Zhou Zhou³. ¹Ruhr-University Bochum ²University College London ³University of Toronto

3:55 PM Multi-Sequence Segmentation Using Score Tests
Hock Peng Chan. National University of Singapore

4:20 PM Computationally Efficient Methods for Multivariate Change-point Analysis with Industrial Application
Idris Eckley. Lancaster University

4:45 PM DNA Copy Number Profiling Using Single-cell Sequencing
Xuefeng Wang. H. Lee Moffitt Cancer Center & Research Institute

5:10 PM Floor Discussion.

Session 111: The challenges of non-constant hazard ratio: delayed treatment effect, treatment dilution and treatment crossover (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Luo, Xiaodong; Sanofi.
Chair: Quan, Hui; Sanofi.

3:30 PM Change Point Analysis of Correlation in Non-stationary Time Series
Holger Dette¹, Weichi Wu² and Zhou Zhou³. ¹Ruhr-University Bochum ²University College London ³University of Toronto

3:55 PM Multi-Sequence Segmentation Using Score Tests
Hock Peng Chan. National University of Singapore

4:20 PM Computationally Efficient Methods for Multivariate Change-point Analysis with Industrial Application
Idris Eckley. Lancaster University

4:45 PM DNA Copy Number Profiling Using Single-cell Sequencing
Xuefeng Wang. H. Lee Moffitt Cancer Center & Research Institute

5:10 PM Floor Discussion.

Session 111: The challenges of non-constant hazard ratio: delayed treatment effect, treatment dilution and treatment crossover (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Luo, Xiaodong; Sanofi.
Chair: Quan, Hui; Sanofi.

3:30 PM Change Point Analysis of Correlation in Non-stationary Time Series
Holger Dette¹, Weichi Wu² and Zhou Zhou³. ¹Ruhr-University Bochum ²University College London ³University of Toronto

3:55 PM Multi-Sequence Segmentation Using Score Tests
Hock Peng Chan. National University of Singapore

4:20 PM Computationally Efficient Methods for Multivariate Change-point Analysis with Industrial Application
Idris Eckley. Lancaster University

4:45 PM DNA Copy Number Profiling Using Single-cell Sequencing
Xuefeng Wang. H. Lee Moffitt Cancer Center & Research Institute

5:10 PM Floor Discussion.

Session 111: The challenges of non-constant hazard ratio: delayed treatment effect, treatment dilution and treatment crossover (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Luo, Xiaodong; Sanofi.
Chair: Quan, Hui; Sanofi.

3:30 PM Change Point Analysis of Correlation in Non-stationary Time Series
Holger Dette¹, Weichi Wu² and Zhou Zhou³. ¹Ruhr-University Bochum ²University College London ³University of Toronto

3:55 PM Multi-Sequence Segmentation Using Score Tests
Hock Peng Chan. National University of Singapore

4:20 PM Computationally Efficient Methods for Multivariate Change-point Analysis with Industrial Application
Idris Eckley. Lancaster University

4:45 PM DNA Copy Number Profiling Using Single-cell Sequencing
Xuefeng Wang. H. Lee Moffitt Cancer Center & Research Institute

5:10 PM Floor Discussion.
4:20 PM Adjusting for Subsequent Therapy in Oncology Trials: A Regression-Based Imputation Method
   • Tommy Fu, Alan Wan and Xiaolong Luo. 1Celgene
   2Celgene

4:45 PM Design and monitoring of survival trials with delayed treatment effect and treatment crossover
   • Xiaodong Luo, Xuezhou Mao, Xian Chen, Junshan Qiu, Steven Bai and Hai Quan. 1Sanofi 2FDA

5:10 PM Floor Discussion.

Session 112: Adaptive Design (Invited)
Room: WILLIFORD A, THIRD FLOOR
Organizer: Tsong, Yi; FDA.
Chair: Qi Xia; Temple University.
3:30 PM Design and Monitoring of Survival Trials with Delayed Treatment Effect and Treatment Crossover
   • Meiyu Shen and Lixin (Leo) Xu. CDER, FDA

3:55 PM Dilemma on Conditional Power-Based Adaptive Sample Size Re-estimation
   • Xiaoyu Cai, Yi Tsong, Meiyu Shen and Yu-Ting Weng.

4:20 PM Utilizing Seamless Adaptive Designs for NASH Clinical Trials
   • Yeh-Fong Chen, Feiran Jiao and George Kordzakhia. US Food and Drug Administration

4:45 PM Floor Discussion.

Session 113: Recent Development and Applications of Statistical Methodology in Rare Diseases (Invited)
Room: CONFERENCE ROOM 4I, FOURTH FLOOR
Organizers: Wu, Jingtao/Qian, Jane; Takeda
Chair: Cao, Charlie; Takeda.
3:30 PM Randomized phase II trials vs. single arm phase II trials with a registry control in rare diseases
   • Brent Logan, Ruta Brazauskas and Raphael Fraser. Medical College of Wisconsin

3:55 PM What constitute Scientific Evidence? The Similarity Principle
   • Mark Chang. Veristat

4:20 PM Adaptive Multi-Stage Clinical Trial Design for Binary Endpoint in Rare Disease
   • Zhaowei Hua and Lingrui Gan. 1Takeda Pharmaceutical International Co. 2University of Illinois at Urbana and Champaign

4:45 PM A Bayesian prediction model between a biomarker and the clinical endpoint for dichotomous variables
   • Yang Song, Zhwei Jianga, Qiong Shou, Jielai Xia and William Wang. 1Vertex Pharmaceuticals Inc., 2Merck Research Labs, 3Fourth Military Medical University of China

5:10 PM Floor Discussion.

Session 114: Statistical Genetics & Genomics (or Computational Biology) (Invited)
Room: CONFERENCE ROOM 4H, FOURTH FLOOR
Organizer: Wu, Song; Stonybrook University.
Chair: Yang Liu; AbbVie Inc.
3:30 PM GEE-based SNP set association testing for longitudinally-measured continuous and discrete traits
   Zuoheng Wang. Yale University

3:55 PM Weighted score method for twin imaging study detects genetic and environmental effects in Human conn
   • Yimei Li, Shenghua Mao and Hongtu Zhu. 1St. Jude Children’s Research Hospital 2MD anderson cancer center

4:20 PM How to define negative data in tumor heterogeneity study?
   • Sujana Ganaparit, Mitsuko Murakami, Li Yan, Jian-min Wang, Changxing Ma, Song Liu and • Lei Wei. 1Roswell Park Cancer Institute 2University at Buffalo

4:45 PM Integrated Statistical Inference on Genomic Handles of Traits (InSIGHT)
   • Cheng Cheng and Lei Shi. St. Jude Children’s Research Hospital

5:10 PM Floor Discussion.

Session 115: Recent Advances in Dose Finding studies (Invited)
Room: CONFERENCE ROOM 4P, FOURTH FLOOR
Organizer: Xi, Dong; Novartis.
Chair: Tu Xu; AbbVie Inc.
3:30 PM Quantitative consideration in dose-finding study designs
   • Hulin Hu and Dong Xi. Novartis Pharmaceutical Corp

3:55 PM Dose Response Models for Longitudinal Data
   CHYIHUNG HSU. Janssen Research & Development, Raritan, NJ, USA

4:20 PM Assessing the similarity of dose response and target doses in two non-overlapping subgroups
   Frank Bretz. Novartis

4:45 PM Discussant: Lei Nie

5:10 PM Floor Discussion.

Session 116: Partial Identification: Seeking reasoned evidence from difficult data (Invited)
Room: CONFERENCE ROOM 4R, FOURTH FLOOR
Organizer: Xia, Michelle; AbbVie Inc.
Chair: Haiming Zhou; Northern Illinois University.
3:30 PM Bayesian inference for partially identified models: Exploring the limits of limited data
   Paul Gustafson. University of British Columbia

3:55 PM Estimating the prevalence of accounting misconduct: A semiparametric Bayesian approach
   Richard Hahn, Jared Murray and Ioanna Manolopoulou. 1University of Chicago 2Carnegie Mellon University 3University College of London

4:20 PM Partially Identified Treatment Effects for Generalizability
   Wendy Chan. University of Pennsylvania

4:45 PM Discussant: Yeh-fong Chen

5:10 PM Floor Discussion.
Session 117: High dimensional multivariate analysis and its applications (Invited)
Room: CONFERENCE ROOM 5G, FIFTH FLOOR
Organizer: Zhong, Ping-Shou; Michigan State University.
Chair: Zhong, Ping-Shou; Michigan State University.
3:30 PM Two-Sample Tests for Sparse High Dimensional Multinomial Distributions
Amanda Plunkett1 and Juyoung Park2. 1Department of Defense, U.S.A. 2University of Maryland Baltimore County
3:55 PM Change-point detection for locally dependent data
Hao Chen. University of California, Davis
4:20 PM Test for the mean matrix in a Growth Curve model for high dimensions
Muni Srivastava. University of Toronto
4:45 PM New Insights in High Dimensional Tests with Applications to Genetic and Genomic Studies
Yaozhou Xue1 and Danning Li2. 1Penn State University 2Jilin University
5:10 PM Floor Discussion.

Session 118: Recent Developments in Graphical Models and Network Analysis (Invited)
Room: CONFERENCE ROOM 5H, FIFTH FLOOR
Organizer: Zhu, Yunzhang; Ohio State University.
Chair: Xuan Bi; Yale University.
3:30 PM Low-rank Tensor Recovery via Cubic-Sketching
Botao Hao1, Anru Zhang2 and Guang Cheng1. 1Purdue University 2University of Wisconsin Madison
3:55 PM Estimation of Gaussian Graphical Model from Data with Dependent Noise Structure
Yuying Xie1, Yufeng Liu2 and William Valdar2. 1Michigan State University 2University of North Carolina at Chapel Hill
4:20 PM Nonparametric Seeded Network Matching
Yuan Zhang. The Ohio State University
4:45 PM Floor Discussion.

Session 119: Novel statistical methods for genetic data analysis (Invited)
Room: CONFERENCE ROOM 5F, FIFTH FLOOR
Organizer: Cui, Yuehua; Michigan State University.
Chair: He, Tao.
3:30 PM Empirical Likelihood Ratio Tests for Coefficients in High Dimensional Heteroscedastic Linear Models
Hongfang Wang1, Ping-Shou Zhong2 and Yuehua Cui2. 1Indiana University-Purdue University Indianapolis 2Michigan State University
3:55 PM Improving Rare Variant Calling with Read Information from Next Generation Sequencing Data
Kai Zhang1 and Degui Li2. 1Michigan Technological University 2The University of Texas Health Science Center
4:20 PM Secondary traits - rare variants association analyses in case-control sequencing studies
Guolian Kang. St. Jude Children’s Research Hospital
4:45 PM Random Field Modelling of Genetic Association in the presence of disease heterogeneity
Ming Li1, Zihuai He2 and Qing Lu3. 1Indiana University at Bloomington 2University of Michigan 3Michigan State University
5:10 PM Floor Discussion.

Session 120: Complex Data Analysis: Methodologies and Applications (Invited)
Room: BOULEVARD AB ROOM, SECOND FLOOR
Organizer: Hu, X. Joan; Simon Fraser University.
Chair: Xiong, Yi; Simon Fraser University.
3:30 PM Bayesian sensitivity analysis for unmeasured confounding in causal mediation analysis
Lawrence McCandless. Simon Fraser University
3:55 PM Functional Principal Component Analysis for Longitudinal and Survival Data
Liangliang Wang and Janghu Dong. Simon Fraser University
4:20 PM Can we use machine learning with large administrative datasets? Should we?
Robert Platt. McGill University
4:45 PM Analysis of Longitudinal Data with Missing Observations or Measurement Error
Grace Yi1, Xianming Tang2 and Runze Li3. 1University of Waterloo 2UNC Chapel Hill 3Penn State
5:10 PM Floor Discussion.

Session 121: Multiplicity in Clinical Trials (Invited)
Room: WILLFORD B, THIRD FLOOR
Organizer: Xi, Dong; Novartis.
Chair: Xi, Dong; Novartis.
3:30 PM Multiple endpoints in clinical trials’ the regulatory, statistical, and common sense perspectives
Dror Rom and Jaclyn McGaugh. Prosoft Clinical
3:55 PM A Gatekeeping Test in a Group Sequential Design with Multiple Interim Looks
Ajit Tamhane1, Jiagao Gou2, Christopher Jennison3, Cyrus Mehta1 and Teresa Corto4. 1Northwestern University 2Hunter College of CUNY 3University of Bath 4Cytel Inc.
4:20 PM Statistical Considerations in Un-Blinded Sample Size Re-Estimation in a Phase 3 Trial
Xin Wang, Sheng Zhong, Yijie Zhou and Lu Cui. Abbvie
4:45 PM Discussant: John Lawrence
5:10 PM Floor Discussion.

Session 122: high dimensional modeling in medicine (Invited)
Room: CONFERENCE ROOM 5I, FIFTH FLOOR
Organizer: Li, Zhigang; Dartmouth.
Chair: Zhao, Lihui; Northwestern University.
Session 122: Recent experience in accelerated approval in Oncology Immunotherapy from Industrial and FDA perspective (Invited)
Room: CONFERENCE ROOM 1Q, FOURTH FLOOR
Organizer: Tang, Rui (Sammi); Vertex Pharmaceuticals.
Chair: Tang, Rui (Sammi); Vertex Pharmaceuticals.
3:30 PM Accelerated Approval of Keytruda in Advanced Melanoma: The Pain and the Glory
Nicole (Xiaoyun) Li and Cong Chen. Merck & Co.
3:55 PM Recent experience in accelerated approval in Oncology from FDA perspective
Vivian Yuan, Kun He and Rajeshwari Sridhara. FDA
4:20 PM Discussant: Huang, Bo
4:45 PM Floor Discussion.

Session 123: Utilizing Real World Evidence in Regulatory Decision: Statistical Considerations and Beyond (Invited)
Room: CONFERENCE ROOM 5E, FIFTH FLOOR
Organizer: Lu, Nelson; FDA.
Chair: Yan, Xu; FDA.
3:30 PM The Use of Real-World Evidence for Making Regulatory Decisions
Greg Campbell. GCStat Consulting
3:55 PM Incorporating Real World Evidence for Regulatory Decision Making with Propensity Score Methodology
Lilly Yue. U.S. FDA
4:20 PM Roadmap from Real World Data to Real World Evidence: Regulatory Experience with Orthopedic Devices
Jianxiang Chu. Food and Drug Administration (FDA)
4:45 PM Use of Real World Evidence in Cardiovascular Device Studies: A Statistical Reviewer’s Experience
Heng Li and Vandana Mukhi. CDRH/Food and Drug Administration
5:10 PM Floor Discussion.

Session 124: Uncertainty, Effect Size and Bias in Statistical Evidence (Invited)
Room: CONFERENCE ROOM 5A, FIFTH FLOOR
Organizer: Liu, Qing; QRMedSci.
Chair: Glen Laird; Vertex Pharmaceuticals.
3:30 PM P-values, bias and variability: a solution for replication power
Laura C. Lazzeroni. Professor of Psychiatry and, by courtesy, of Medicine and Biomedical Data Science, Stanford University
3:55 PM Fair prediction with disparate impact: A study of bias in recidivism prediction instruments
Alexandra Chouldechova. Heinz College, Carnegie Mellon University
4:20 PM Beyond p-values: a phase II design with statistical significance and clinical relevance
Satrajit Roychoudhury. Novartis Pharmaceutical Corporation
4:45 PM Floor Discussion.

Session 125: Topics on Clinical Statistics (Contributed)
Room: CONFERENCE ROOM 5B, FIFTH FLOOR
Chair: Rama Melkote; Johnson and Johnson.
3:30 PM Device Trials: Study Design Considerations
Rama Melkote. Janssen R&D
3:45 PM Benefit and Risk Assessment in Ophthalmic Device Clinical Studies
Chul Ahn. FDA
4:00 PM Method Comparison for Diagnostic Devices that Measure Different Parameters
Changhong Song. Food and Drug Administration
4:15 PM Some Considerations for the Trade-off Assessment of Diagnostic Errors when Comparing Diagnostic Test
Norberto Pantoja-Galicia and Gene Pennello. US Food and Drug Administration/CDRH
4:30 PM Readmission Time Analysis for Psychiatry Patients Based on a Cox Intensity Process Model for Recurrence
Qing Li and Gideon Zamba. The University of Iowa
4:45 PM Rank Selection for Multilinear Principal Component Analysis
I-Ping Tu1, Su-Yun Huang1 and Dai-Ni Hsieh2. 1Institute of Statistical Science, Academia Sinica
5:00 PM Floor Discussion.

Session 126: Topics on Applied Statistics (II) (Contributed)
Room: WILLIFORD C, THIRD FLOOR
Chair: Chenghao Chu; Indiana Purdue University.
3:30 PM Assessing skeletal growth rate around pubertal growth spurt
Chenghao Chu, Ying Zhang and Wanzhu Tu. Department of Biostatistics, IUPUI
3:45 PM Benefit and Risk Assessment in Ophthalmic Device Clinical Studies
Chul Ahn. FDA
4:00 PM Method Comparison for Diagnostic Devices that Measure Different Parameters
Changhong Song. Food and Drug Administration
4:15 PM Some Considerations for the Trade-off Assessment of Diagnostic Errors when Comparing Diagnostic Test
Norberto Pantoja-Galicia and Gene Pennello. US Food and Drug Administration/CDRH
4:30 PM Readmission Time Analysis for Psychiatry Patients Based on a Cox Intensity Process Model for Recurrence
Qing Li and Gideon Zamba. The University of Iowa
4:45 PM Rank Selection for Multilinear Principal Component Analysis
I-Ping Tu1, Su-Yun Huang1 and Dai-Ni Hsieh2. 1Institute of Statistical Science, Academia Sinica
5:00 PM Floor Discussion.
Wednesday, June 28. 8:30 AM - 10:10 AM

Session 128: Statistical Analysis in High-Dimensional Data Analysis with Applications (Invited)
Room: CONFERENCE ROOM 4F, FOURTH FLOOR
Organizer: Li, Dongmei; Rochester University.
Chair: Li, Dongmei; Rochester University.

8:30 AM An ensemble method for RNA-Seq differential analysis
♦ Dongmei Li, Martin Zand, Thomas Fogg, Zidian Xie and Timothy Dye. University of Rochester

8:55 AM Association tests for sparse genetic data in structured and correlated samples
♦ Han Chen and Xi Hong Lin. 1Harvard T.H. Chan School of Public Health

9:20 AM Confident Inference for SNP Effects on Treatment Efficacy
♦ Ying Ding 2 and Jason Hsu 1. 1University of Pittsburgh 2Ohio State University

9:45 AM Multiple Imputation for Missing Data in Analgesic Clinical Trials using Pattern Mixture Models
♦ Xueya Cai 1, Michael McDermott 1, Jennifer Gewandter 1, Hu He 2 and Robert Dworkin 1. 1University of Rochester 2Tulane University

10:10 AM Floor Discussion.

Session 129: Biomarker Integration in Cancer Clinical Trials (Invited)
Room: CONFERENCE ROOM 4H, FOURTH FLOOR
Organizer: Polley, Mei-Yin; Mayo Clinic.
Chair: Lu, Ying; Stanford University.

8:30 AM Statistical Issues in Biomarker Clinical Trials
Sumithra Mandrekar. Mayo Clinic

8:55 AM An adaptive design for the identification of the optimal dose using joint modelling of biomarker measurements and toxicity over all treatment cycles in phase I/II clinical trials of molecularly targeted agents in oncology
♦ Maria-Athina Alzgerinakou 1 and Xavier Pauletti 2. 1CESP OncoStat, INSERM, University of Paris-Saclay 2Institut Gustave Roussy, CESP OncoStat

9:20 AM Efficient Basket Trial Designs
♦ Kristen Cunanan, Alexia Iasonos, Ronglai Shen, Colin Begg and Mithat Gonen. Memorial Sloan Kettering Cancer Center

9:45 AM Two-Stage Adaptive Cutoff Design for Building and Validating a Prognostic Biomarker Signature
♦ Mei-Yin Polley 1, Eric Polley 1, Erich Huang 2, Boris Freidlin 2 and Richard Simon 2. 1Mayo Clinic 2National Cancer Institute

10:10 AM Floor Discussion.

Session 130: Statistical considerations in equivalence assessment of patch products using data of adhesive and irritation study (Invited)
Room: WILLIFORD A, THIRD FLOOR
Organizer: Tsong, Yi; FDA.
Chair: Mengdie Yuan; FDA.

8:30 AM Statistical Considerations in CDER’s New Draft Guidance for Assessing Adhesion with Transdermal Delivery
♦ Wanjie Sun and Stella Grosser. FDA

8:55 AM Statistical Considerations of Adhesion Data Analysis
♦ Yu-Ting Weng, Caroline Strasinger and Meiya Shen. FDA

9:20 AM Statistical Issues in Irritation/Sensitization Study for Transdermal Drug Products
♦ Jingyu (Julia) Luan, Mengdie Yuan and Stella Grosser. US FDA

9:45 AM Statistical Analysis and Issues in Skin Adhesion Studies
Pina D'Angelo. Novum Pharmaceutical Research Services

10:10 AM Floor Discussion.

Session 131: Recent Advances in Data Analysis with Special Structures (Invited)
Room: CONFERENCE ROOM 4P, FOURTH FLOOR
Organizer: Yang, Dan; Rutgers University.
Chair: Yang, Dan; Rutgers University.

8:30 AM Bayesian Multiple Change Point Detection with Non-local Priors
♦ Fei Jiang 1, Guoshen Yin 1 and Francesca Dominici 2. 1The University of Hong Kong 2Harvard University

8:55 AM Slope meets Lasso in sparse linear regression
♦ Pierre C Bellec 1, Guillaume Lecue 1 and Alexandre B Tsybakov 3. 1Rutgers 2ENSAM-CREST

9:20 AM Nonparametric Methods in Business Analytics
Haiping Shen. University of Hong Kong

9:45 AM Factor Models for Matrix-Valued High-Dimensional Time Series
♦ Dong Wang 1, Xialu Liu 2 and Rong Chen 3. 1Princeton University 2San Diego State University 3Rutgers University

10:10 AM Floor Discussion.

Session 132: High dimensional inference and its application to change-point analysis (Invited)
Room: CONFERENCE ROOM 4R, FOURTH FLOOR
Organizer: Zhong, Ping-Shou; Michigan State University.
Chair: Zhong, Ping-Shou; Michigan State University.
8:30 AM High-dimensional covariance matrix estimation
  *Yilei Wu, Yingli Qin and Mu Zhu. University of Waterloo

8:55 AM Simultaneous Inference for Multiple Change points
  Ning Hao. University of Arizona

9:20 AM Test for temporal homogeneity of high-dimensional means with application to fMRI studies
  *Jun Li and Ping-Shou Zhong2. 1Kent State University
  2Michigan State University

9:45 AM Sparse Multivariate Statistics with Discrete Optimization
  *Rahul Mazumder. Massachusetts Institute of Technology

10:10 AM Floor Discussion.

Session 133: Leadership and Career Development (Invited)
Room: CONFERENCE ROOM 5G, FIFTH FLOOR
Organizer: Zhou, Yijie; AbbVie Inc.
Chair: Zhou, Yijie; AbbVie Inc.
8:30 AM Statistical Leadership and Career Development in Pharmaceutical Industry
  Ivan Chan. AbbVie, Inc.

8:55 AM Career Development for Biostatisticians in Academia
  Haitao Chu. University of Minnesota

9:20 AM My Leadership Experience in Government
  Yi Tsong. CDER, FDA

9:45 AM Discussant: Bo Yang

10:10 AM Floor Discussion.

Session 134: Recent Advances in Analytical Methods for Cancer Genomics (Invited)
Room: CONFERENCE ROOM 5I, FIFTH FLOOR
Organizer: Zhu, Bin; NIH.
Chair: Zhu, Bin; NIH.
8:30 AM Probabilistic models and statistical methods in cancer etiology and evolution
  Cristian Tomasetti. Johns Hopkins University

8:55 AM Quantifying tumor evolution via spatial computational modeling and Approximate Bayesian Computing
  Christina Curtis. Stanford University School of Medicine

9:20 AM MEGSA: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations
  *Xing Hua, Pauala Hyland, Jing Huang, Bin Zhu, Neil Caporaso, Maria Landi and Jianxin Shi. National Cancer Institute

9:45 AM Floor Discussion.

Session 135: Recent developments in statistical machine learning (Invited)
Room: CONFERENCE ROOM 5E, FIFTH FLOOR
Organizer: Zhu, Yunzhang; Ohio State University.
Chair: Zhang, Yuan; Ohio State University.
8:30 AM Multilayer tensor factorization with applications to recommender systems
  *Xuan Bi1, Annie Qu2 and Xiaotong Shen3. 1Yale University 2University of Illinois at Urbana-Champaign
  3University of Minnesota Twin Cities

8:55 AM A new SVD approach to optimal topic estimation
  *Tracy Ke and Minche Wang. University of Chicago

9:20 AM On the connections between algorithmic regularization and penalization for GLM's
  *Qian Qian, Vincent Vu and Yunzhang Zhu. The Ohio State University

9:45 AM Floor Discussion.

Session 136: Recent advances in the analysis of complex data (Invited)
Room: WILLIFORD C, THIRD FLOOR
Organizer: Wu, Yichao; NC State University.
Chair: Wu, Yichao; NC State University.
8:30 AM Shape Constrained Tensor Factorizations
  *Eric Chi1, Bethany Lusch2 and Nathan Kutz2. 1North Carolina State University 2University of Washington

8:55 AM Matrix Linear Discriminant Analysis
  Wei Hu1, Weining Shen1, Hua Zhou2 and *Dehan Kong3. 1University of California, Irvine 2University of California, Los Angeles 3University of Toronto

9:20 AM A Note on Inverse Regressions When Responses are Missing at Random
  Yaexiao Dong, Cheng Yong Tang and *Qi Xia. Temple University

9:45 AM Nonparametric estimation of multivariate mixture
  *Chaowen Zheng and Yichao Wu. North Carolina State University

10:10 AM Floor Discussion.

Session 137: Statistical Inference for High-dimensional Linear Regression and Covariance Structure (Invited)
Room: CONFERENCE ROOM 4J, FOURTH FLOOR
Organizer: Zhang, Anru; University of Wisconsin.
Chair: Zhang, Anru; University of Wisconsin.
8:30 AM Structured Volatility Matrix Estimation with Accurate Covolatility Estimators based on Non-synchron
  Jingqing Fan, *Donggyu Kim and Kun Lu. Princeton University

8:55 AM Selection of Effective Scores for Treatment Effect Estimation
  *Lei Wang, Ying Zhang, Menggang Yu and Jun Shao. University of Wisconsin-Madison

9:20 AM Structured Correlation Detection with Application to Colocalization Analysis in Microscopic Imaging
  *Shulei Wang1, Jianqing Fan2, Ginger Pocock1 and Ming Yuan1. 1University of Wisconsin-Madison 2Princeton University

9:45 AM CHIME: Clustering of High-Dimensional Gaussian Mixtures with EM Algorithm and Its Optimality
  Tony Cai, Jing Ma and *Linjun Zhang. University of Pennsylvania

10:10 AM Floor Discussion.
**Session 138: Recent advances in statistical genomics (Invited)**  
Room: ASTORIA ROOM, THIRD FLOOR  
Organizer: Zhao, Hongyu; Yale University.  
Chair: Zhao, Hongyu; Yale University.

8:30 AM Practical Issues of meta-analysis of genetic associations using sequence data  
♦ Yu Jiang. Penn State College of Medicine

8:55 AM A graphical model to prioritizing GWAS results guided by biomedical literature mining  
♦ Dongjun Chang¹, Hang Kim², Zhenning Yu³, Andrew Lawson¹ and Hongyu Zhao³. ¹Medical University of South Carolina ²University of Cincinnati ³Yale University

9:20 AM Structured subcomposition selection in regression and its application to microbiome data analysis  
Tao Wang. Shanghai Jiao Tong University

9:45 AM AC-PCA: simultaneous dimension reduction and adjustment for confounding variation  
♦ Zhixiang Lin¹, Can Yang², Hongyu Zhao³ and Wing Hung Wong¹. ¹Stanford University ²Hong Kong Baptist University ³Yale University

10:10 AM Floor Discussion.

**Session 139: Analysis of High-Dimensional Genomics Data (Invited)**  
Room: CONFERENCE ROOM 4Q, FOURTH FLOOR  
Organizer: Cai, Tony; University of Pennsylvania.  
Chair: Cai, Tony; University of Pennsylvania.

8:30 AM Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach  
Dangeng Liu¹ and Heping Zhang². ¹University of Cincinnati Lindner College of Business ²Yale University School of Public Health

8:55 AM Non-parametric Empirical Bayes method for sparse, noisy signals  
Junhui Cai¹, Ya’acov Ritov² and Linda Zhao¹. ¹University of Pennsylvania ²University of Michigan

9:20 AM False Discovery Rate Control for High-Dimensional Networks of Conditional Quantile Associations  
♦ Jichun Xie¹ and Ruoshua Li². ¹Duke University ²The University of Texas at Houston

9:45 AM Floor Discussion.

**Session 140: Recent Development of Statistical Learning Methods for Precision Medicine (Invited)**  
Room: CONFERENCE ROOM 4B, FOURTH FLOOR  
Organizer: Lu, Wenbin; NC State University.  
Chair: Song, Rui; NC State University.

8:30 AM Subgroup Identification from Multiple Individual Patient Level Data Sets  
♦ Chensheng Kuang¹, Menggeng Yu¹, Qi Tang², Alan Hartford² and Weining Robieson². ¹University of Wisconsin-Madison ²Sanofi ²AbbVie

8:55 AM VARIABLE SELECTION FOR ESTIMATING THE OPTIMAL TREATMENT REGIMES  
Baqun Zhang¹ and Min Zhang². ¹Shanghai University of Finance and Economics ²University of Michigan

9:20 AM Inference for Two-Stage Dynamic Treatment Regimes in the Presence of Drop-Out  
♦ Abdas Wahed³ and Andrew Topp². ²Dept. of Biostatistics, University of Pittsburgh ³AbbVie

9:45 AM Statistical Learning Methods to Improve Risk Prediction in Organ Transplantation  
Lihai Zhao. Northwestern University

10:10 AM Floor Discussion.

**Session 141: New developments and Challenges in Biomarker Discovery (Top Contributed)**  
Room: CONFERENCE ROOM 5B, FIFTH FLOOR  
Organizer: Veavi Chang; Eli Lilly.  
Chair: Veavi Chang; Eli Lilly.

8:30 AM Protein Biomarker Discovery from ELISA Cytokines Multiplex Assays  
Veavi Chang. Eli Lilly and Company

8:50 AM Detection for the Onset of Alzheimer’s Disease Using Joint Modeling  
♦ Tian Wang and Jimin Ding. Washington University in St. Louis

9:10 AM Scalable Analysis for Mass Spectrometry Imaging Experiments  
♦ Kylie Bemis and Olga Vitek. Northeastern University

9:30 AM Enhancement of the Adaptive Signature Design for Learning and Confirming in a Single Pivotal Trial  
Gu Mi. Eli Lilly and Company

9:50 AM A non-randomized procedure for discrete multiple testing based on randomized tests  
♦ XIAYU DAI, Nan Lin, Daofeng Li and Ting Wang. Washington University in St. Louis

10:10 AM Floor Discussion.

**Session 142: Topics on Applied Statistics (III) (Contributed)**  
Room: CONFERENCE ROOM 5F, FIFTH FLOOR  
Chair: Jay Mandrekar; Mayo Clinic.

8:30 AM Optimal Number of Tissue Specimens Required for Diagnosis of Prosthetic Joint Infection  
Jay Mandrekar. Mayo Clinic

8:45 AM Higher moments modified VaR estimators and their applications in portfolio  
Shu-Hui Yu. National University of Kaohsiung

9:00 AM Permutation Test Based on a Linear Mixed Model for Differential Alternative Splicing Analysis  
Huining Kang. University of New Mexico

9:15 AM Finding observed Fisher information matrix for EM based inference on misrepresentation risk in insur  
♦ Restof Akakpo¹, Michelle Xia¹ and Alan Polansky. ¹Northern Illinois University
Wednesday, June 28. 10:30 AM - 12:10 PM

Session 144: Recent Advances in Quantile Regression (Invited)
Room: CONFERENCE ROOM 4I, FOURTH FLOOR
Organizer: Zhang, Jing; University of Maryland.
Chair: Zhang, Jing; University of Maryland.
8:30 AM Hierarchical Models for Combining N-of-1 Trials
Christopher Schmid, Brown University
8:55 AM Multiple dosage levels in network meta-analysis using the potential outcomes approach: A comparison
• Mireille Schnitzer1 and Michele Bally2. 1Université de Montréal 2Centre hospitalier l’Université de Montréal
9:20 AM The impact of excluding trials from network meta-analyses: an empirical study
• Jing Zhang1, Yiping Yuan2 and Haitao Chu2. 1University of Maryland 2University of Minnesota
9:45 AM A Bayesian meta-analysis method for estimating risk difference of rare events
Yaonvan Tang, Saint Luke’s Health System
10:10 AM Floor Discussion.

Wednesday, June 28. 10:30 AM - 12:10 PM

Session 145: Recent Developments in Statistical Methods in Healthcare (Invited)
Room: CONFERENCE ROOM 4B, FOURTH FLOOR
Organizer: Wang, Dong; Princeton University.
Chair: Wang, Dong; Princeton University.
10:30 AM Precision Care of Stroke Patients: Trajectory Analysis of Dynamic Blood Pressure
Hai Zhi, University of Hong Kong
10:55 AM Tree based weighted learning for estimating individualized treatment rules with censored data
• Yifan Cui1, Ruqing Zhu2 and Michael Kosorok1. 1University of North Carolina at Chapel Hill 2University of Illinois at Urbana-Champaign
11:20 AM On Reject and Refine Options in Multicategory Classification
Chong Zhang1, Wenbo Wang2 and Xingye Qiao2. 1Seattle WA 2Binghamton University
11:45 AM Predicting Hospital Admissions from Emergency Departments at Triage
Han Ye1, Zhankun Sun2 and Haipeng Shen3. 1University of Illinois at Urbana Champaign 2City University of Hong Kong 3University of Hong Kong
12:10 PM Floor Discussion.

Session 146: Recent Advances on Neuroimaging Data Analysis (Invited)
Room: CONFERENCE ROOM 5F, FIFTH FLOOR
Organizer: Wang, Lily; Iowa State University.
Chair: Wang, Lily; Iowa State University.
10:30 AM Optimal experimental designs for fMRI when the model matrix is uncertain
• Ming-Hung Kao and Lin Zhou. Arizona State University
10:55 AM Simultaneous Confidence Bands for Mean and Variance Functions based on Deterministic Design
Li Cai1, Rong Liu2, Suojin Wang3 and Lijian Yang4. 1University of Hong Kong 2University of Toledo 3Texas A&M University 4Tsinghua University
11:20 AM Spatially Varying Coefficient Models
• Jingru Mu1, Guannan Wang2 and Lily Wang1. 1Iowa State University 2College of William & Mary
11:45 AM Functional data analysis for dynamic PET data
• R. Todd Ogden, Yakanun Chen and Jeff Goldsmith. Columbia University, Department of Biostatistics, New York, NY
12:10 PM Floor Discussion.

Session 147: Statistical learning for complex data (Invited)
Room: CONFERENCE ROOM 4I, FOURTH FLOOR
Organizer: Lu, Henry Horng-Shing; National Chiao Tung University Taiwan.
Chair: Lu, Henry Horng-Shing; National Chiao Tung University Taiwan.
10:30 AM High-dimensional Linear Regression for Dependent Data with Applications to Nowcasting
• Yuefeng Han and Ruye Tsay. University of Chicago
10:55 AM Model selection for high-dimensional sparse nonlinear models using Chebyshev greedy algorithms
Ching-Kang Ing1, You-Lin Chen2 and Wei Biao Wu2. 1National Tsing Hua University 2University of Chicago
11:20 AM Classical Backfitting for Smooth-backfitting Additive Models
• Li-Shan Huang and Chung-Hsin Yu. National Tsing Hua University, TAIWAN
11:45 AM Online Learning with Applications to Communication Network Traffic Management

*Henry Horng-Shing Lu.* National Chiao Tung University, Taiwan

12:10 PM Floor Discussion.

**Session 148: Recent Developments in Theory and Application of Multiple Comparison Methods (Invited)**

Room: CONFERENCE ROOM 4F, FOURTH FLOOR
Organizer: Qian, Jane; AbbVie Inc.
Chair: Qian, Jane; AbbVie Inc.

10:30 AM Partition to Power Subgroup Inference in Personalized Medicine

*Szu-Yu Tang*¹, *Yi Liu*² and *Jason Hsu*³. ¹Ventana Medical Systems, Inc. ²Takeda Pharmaceuticals International Co. ³The Ohio State University

10:55 AM Multiplicity issues in clinical trials from a regulatory perspective

*Yongnan Kim.* FDA

11:20 AM Lossless type-I error rate control for multiplicity and its application

*Muhammad Jalaluddin.* AbbVie

11:45 AM A Gatekeeping Test on a Primary and a Secondary Endpoint in a Group Sequential Design with Multiple

*Ajit Tamhane*¹, *Jiangtao Gou*², *Christopher Jennison*³, *Cyrus Mehta*² and *Teresa Curto*³. ¹Northwestern University ²Hunter College ³University of Bath ⁴Cytel, Inc.

12:10 PM Floor Discussion.

**Session 149: Go/No-Go decision making in clinical trials (Invited)**

Room: CONFERENCE ROOM 4P, FOURTH FLOOR
Organizer: Wang, Deli; AbbVie Inc.
Chair: Wang, Deli; AbbVie Inc.

10:30 AM Enhancing the Probability of Success Framework for Go/No-Go Decision Making Using ROC Curves

*Qi Tang*¹, *Alan Hartford*², *Deli Wang*³, *Dey Jiotirmoy*³, *Walt Offen*⁴ and *Frank Shen*⁵. ¹Translational Informatics, Sanofi, Bridgewater, NJ ²Statistical Innovation, AbbVie, Chicago, IL ³Pipeline Statistics, AbbVie, Chicago, IL ⁴Statistical Sciences, AbbVie, Chicago, IL ⁵Data and Statistical Sciences, AbbVie, Chicago, IL

10:55 AM Interim Go/No-Go decision making in a phase 2 clinical trial of a neurodegenerative disorder

*Weining Robieson, Greg Cicconetti*¹ and *Deli Wang*¹. ¹AbbVie

11:20 AM Statistical Software for Decision Making in Clinical Development

*Charles Liu.* Cytel

11:45 AM Discussant: Marty King

12:10 PM Floor Discussion.

**Session 150: Modern Advancements in Semi-/Non-parametric Statistics (Invited)**

Room: CONFERENCE ROOM 4R, FOURTH FLOOR
Organizer: Wang, Guannan; William and Mary.
Chair: Wang, Guannan; William and Mary.

10:30 AM Free-knot spline for Generalized Regression Models

*Jing Wang.* University of Illinois at Chicago

10:55 AM Sparse Model Identification and Learning for Ultra-high-dimensional Additive Partially Linear Models

*Xinyi Li*¹, *Li Wang*² and *Dan Nettleton*. ¹Iowa State University

11:20 AM Partially time-varying coefficient proportional hazards models with error-prone covariates

*Xiao Song*¹ and *Li Wang*². ¹University of Georgia ²Iowa State University

11:45 AM Floor Discussion.

**Session 151: Modern Statistical methods for Educational Testing (Invited)**

Room: CONFERENCE ROOM 5G, FIFTH FLOOR
Organizer: Wang, Xiaojing; University of Connecticut.
Chair: Wang, Xiaojing; University of Connecticut.

10:30 AM Bayesian estimation of the DINA Q matrix

*Yinghan Chen and Steve Culpepper.* University of Illinois at Urbana-Champaign

10:55 AM Latent Class Models for Learning

*Shiyu Wang, Steve Culpepper, Yinghan Chen and Jeff Douglas.*

11:20 AM A Fused Latent and Graphical Model

*Jingchen Liu.* Columbia University

11:45 AM Practical Considerations for Educational Testing

*Betsy McCoach.* University of Connecticut

12:10 PM Floor Discussion.

**Session 152: Variable Selection and Dimension Reduction for High-Dimensional Models (Invited)**

Room: CONFERENCE ROOM 5H, FIFTH FLOOR
Organizer: Zambom, Adriano Zanin; Loyola University Chicago.
Chair: Zambom, Adriano Zanin; University of Cincinnati.

10:30 AM Linear and Nonlinear Sufficient Dimension Reduction for Functional Data

*Bing Li*¹ and *Jun Song*². ¹Penn State University ²Penn State University

10:55 AM Joint mean and covariance estimation based on model selection with unreplicated matrix-variate data

*Michael Hornstein, Roger Fan, Kerby Shedden and Shuheng Zhou.* University of Michigan

11:20 AM Generalized Principal Component Analysis

*Yoonkyung Lee*¹ and *Andrew Landgraf*². ¹Ohio State University ²Battelle Memorial Institute

11:45 AM Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications

*Xiaohui Chen.* University of Illinois at Urbana-Champaign

12:10 PM Floor Discussion.
Session 153: Advances in joint models for longitudinal and time-to-event data with biomedical applications (Invited)
Room: WILLIFORD A, THIRD FLOOR
Organizer: Zhang, Jingyang; Fred Hutchinson Cancer Research Center.
Chair: Zhang, Jingyang; Fred Hutchinson Cancer Research Center.
10:30 AM Informative Cluster Size and Observation Time in Longitudinal Transition Models
Joe Bible, Danping Liu and Paul Albert. National Institutes of Health
10:55 AM Statistical Methods for Multivariate Failure Time Data Analysis
Shanshan Zhao1 and Ross Prentice2. 1NIH/NIH 2Fred Hutchinson Cancer Research Center
11:20 AM A Frailty-Copula Approach for Modeling a Terminal Event and Recurrent Events with Biomarker Data
Zheng Li1 and Ming Wang2. 1Pennsylvania State University
11:45 AM Joint models of longitudinal outcome and interval-censored events with time varying covariate effect
Bin Zhang1 and Yue Zhang2. 1Cincinnati Children’s Hospital Medical Center 2Shanghai Jiao Tong University
12:10 PM Floor Discussion.

Session 154: Recent Advances on High-Dimensional Statistics (Invited)
Room: CONFERENCE ROOM 5A, FIFTH FLOOR
Organizer: Sun, Wenguang; University of Southern California.
Chair: Sun, Wenguang; University of Southern California.
10:30 AM Simultaneous Multistage Adaptive Ranking and Thresholding for Sparse Signal Recovery
Wenguang Sun and Weinan Wang. University of Southern California
10:55 AM Community Detection in Sparse Networks Using the Symmetrized Laplacian Inverse Matrix (SLIM)
Bingyi Jing, Ting Li, Ningchen Yang and Xianning Yu.
11:20 AM Spatial Adaptation in Trend Filtering
Adityanand Guntuboyina1, Donovan Lieu1, Sabyasachi Chatterjee2 and Bodhisattva Sen3. 1University of California, Berkeley 2University of Chicago 3Columbia University
11:45 AM Floor Discussion.

Session 155: Utilizing External Data in a Medical Product Evaluation Study: Statistical Design and Analysis Consi (Invited)
Room: CONFERENCE ROOM 5I, FIFTH FLOOR
Organizer: Xu, Yuling; FDA.
Chair: Wang, Chenguang; Johns Hopkins University.
10:30 AM Some thoughts in designing a Bayesian study: From a statistical reviewer’s perspective
Xu Yan, Xuan Ye and Yuling Xu. Food and Drug Administration
10:55 AM A Bayesian Non-inferiority Design with Companion Constancy Test in Active Controlled Trials
Ying Yang, Yunling Xu, Ram Tiwari, Nelson Lu and Yu Zhao. Food and Drug Administration
11:20 AM On Weighted Performance Goals in Medical Device Single-Arm Clinical Studies
Nelson Lu, Wei-Chen Chen, Yunling Xu and Ram Tiwari. FDA/CDRH
11:45 AM A Study Design for Utilizing External Data to Augment the Control in a Randomized Controlled Trial
Yunling Xu, Nelson Lu, Lilly Yue and Ram Tiwari. FDA/CDRH
12:10 PM Floor Discussion.

Session 156: Integrative methods for challenging genomics data (Invited)
Room: CONFERENCE ROOM 5E, FIFTH FLOOR
Organizer: Li, Qunhua; Penn State University.
Chair: Lin, Lynn; Penn State University.
10:30 AM Clustering with Hidden Markov Model on Variable Blocks
Lynn Lin. Pennsylvania State University
10:55 AM A hierarchical hidden Markov random field model for peak calling across multiple Hi-C datasets
Joshua Martin1, Ming He2 and Yun Li3. 1University of North Carolina 2Cleveland Clinic Foundation
11:20 AM Bayesian Approaches to Integrative Genomic Analysis: Data Integration and Scalable Computing
Xiaoquan Wen. University of Michigan
11:45 AM Assessing reproducibility of Hi-C chromatin loops using irreproducible discovery rate regression
Feipeng Zhang, Tao Yang and Qunhua Li. Penn State University
12:10 PM Floor Discussion.

Session 157: Functional Data Analysis in Biomedical Research (Invited)
Room: WILLIFORD C, THIRD FLOOR
Organizer: Zhu, Hongxiao; Virginia Tech.
Chair: Hongxiao Zhu; Virginia Tech.
10:30 AM Dynamic Modeling of Longitudinal Snippets
Hans-Georg Mueller and Matthew Dawson. University of California, Davis
10:55 AM Optimal Penalized Function-on-Function Regression under a Reproducing Kernel Hilbert Space Framework
Xiaoxiao Sun1, Pang Da2, Xiao Wang3 and Ping Ma3. 1University of Georgia 2Virginia Tech 3Purdue University
11:20 AM Registration for Binary Functional Data
Julia Wrobel and Jeff Goldsmith. Columbia University
11:45 AM Floor Discussion.

Session 158: Urging a paradigm change: New developments on statistical inferences (Invited)
Room: WILLIFORD B, THIRD FLOOR
Organizer: Xie, Mingye; Rutgers University.
Chair: Jan Hannig; The University of North Carolina at Chapel Hill.
10:30 AM Approximate CD Computing: An effective likelihood-free method with statistical guarantees
*Suzanne Thornton and Min-ge Xie.* Rutgers, The State University of New Jersey

10:55 AM Partial Bayes: Exact Inference with Partially Specified Bayesian Models
*Yixuan Qiu, Lingsong Zhang and Chuanhai Liu.* Purdue University

11:20 AM Personalism and Dempster-Shafer Analysis for the 21st Century
*Paul Edlefsen¹, Ruobin Gong² and Arthur P. Dempster².* ¹Fred Hutchinson Cancer Research Center ²Harvard University

11:45 AM Generalized Fiducial Inference for Nonparametric Function Estimation
*Randy Lai¹ and Raymond Wong².* ¹University of Maine ²Iowa State University

12:10 PM Floor Discussion.

**Session 159: Statistical Modeling of High Dimensional Data with Applications to Genomics (Invited)**

**Room:** ASTORIA ROOM, THIRD FLOOR
**Organizer:** Zhang, Yuping; University of Connecticut.
**Chair:** Zhang, Yuping; University of Connecticut.

10:30 AM Modeling parent-of-origin effect in eQTL mapping using RNA-seq data
*Feifei Xiao¹, Shirong Deng², Guoshuai Cai³ and Christopher I. Amos³.* ¹University of South Carolina ²Augusta University ³Dartmouth College

10:55 AM A Deterministic Global Optimization Method for Variational Inference
*Hachem Saddiki¹, Andrew Trapp² and Patrick Flaherty¹.* ¹University of Massachusetts, Amherst ²Worcester Polytechnic Institute

11:20 AM Statistical framework for 3D chromatin structure modeling from Hi-C data
*Zhengqing Ouyang.* The Jackson Laboratory for Genomic Medicine

11:45 AM Pathway Lasso: Estimate and Select Causal Mediation Pathways with High Dimensional Mediators
*Yi Zhao and Xi Luo.* Brown University

12:10 PM Floor Discussion.
Session 1: Recent progress in high dimensional data analysis

FDR control procedure integrating prior correlation structure with application to microbiome data

Jun Chen¹, Hongyuan Cao² and Jian Xiao¹
¹Mayo Clinic
²University of Missouri

One challenge facing the omics data analysis is the small sample size, compared to the high dimensionality of omics features. Most of the omics studies are underpowered to detect differential features with respect to certain biological and environmental conditions. To boost the statistical power for differential feature discovery, it is critical to integrate the prior structure information into data analysis. Here I propose a resampling-based FDR control procedure that uses the prior correlation structure. The proposed method was motivated by the human microbiome data, where there is a phylogenetic tree relating all the bacteria. The phylogenetic tree provides important evolutionary relationships between bacteria and a prior correlation structure can be defined based on the tree. The prior correlation structure is usually orthogonal to the correlation structure observed in the data. Utilizing the phylogenetic tree information can potentially increase the power of differential abundance analysis since closely related bacterial species tend to exhibit similar biological characteristics and they have a tendency to have the same differential status and/or similar effect sizes. A phylogeny-based FDR control was thus designed to take into account the phylogenetic tree and to increase the power of detecting differentially abundant bacteria. Simulation as well as real microbiome data will be used to illustrate the proposed method.

Robust Factor Models with Covariates

Jianqing Fan¹, Yuan Ke¹ and Yuan Liao²
¹Princeton University
²Rutgers University

We study factor models when the latent factors can be explained partially by several observed covariates. In financial factor models for instance, the unknown factors can be reasonably well predicted by a few covariates, such as the Fama-French factors. To incorporate the explanatory power of these covariates, we propose a two-step estimation procedure: (i) regress the data onto the observables, and (ii) take the principal components of the fitted data to estimate the loadings and factors. With those covariates, the factors can be estimated accurately even if the cross-sectional dimension is mild. The proposed estimator is robust to possibly heavy-tailed distributions, which are encountered in many applications. Empirically, we apply the model to forecasting US bond risk premia, and find that the observed economic covariates contain strong explanatory powers of the factors. The gain of forecast is more substantial when these covariates are incorporated to estimate the common factors than directly used for forecasts.

Variable and Group Selection with Prior Information

Yuan Jiang¹, Li Kai¹, He Yunxiao² and Zhang Heping³
¹Oregon State University
²Nielsen Company
³Yale University

In a multiple regression with high-dimensional predictors, it is usually desired to select important variables or groups of variables. Traditional penalization methods such as lasso and bridge, and their group versions, such as group lasso and group bridge, have been well established for variable selection and group selection. However, in many scientific areas, prior knowledge is available about the importance of predictors or grouped predictors, leading to the necessity of methodological development to incorporate such valuable information. For prior-informative variables and groups of variables, we develop new penalization methods called “prior lasso” and “group ridge”, respectively, to incorporate such prior information into penalized generalized linear models. When the prior information is relatively accurate, both methods were shown to possess theoretical and empirical advantages over their traditional counterparts through asymptotic theories and simulation studies. When the prior information is less reliable, both methods are robust to the misspecification. We also illustrate the applicability and efficacy of both methods using real data sets from genome-wide association studies.

Testing for Trends in High-dimensional Time Series

Likai Chen and Wei Biao Wu
University of Chicago

The paper considers statistical inference for trends of high-dimensional time series. Based on a modified L2-distance between parametric and nonparametric trend estimators, we propose a diagonalized quadratic form test statistic for testing patterns on trends, such as linear, quadratic or parallel forms. We develop an asymptotic theory for the test statistic. A Gaussian multiplier testing procedure is proposed and it has an improved finite sample performance. Our testing procedure is applied to a spatial temporal temperature data gathered from various locations across America. A simulation study is also presented to illustrate the performance of our testing method.

Session 2: Spatial and Temporal Modeling for Environmental Data

Interannual variation in the characteristics of precipitation events under a changing climate

Chen Chen¹, Won Chang², Wenwen Kong³, Jiali Wang⁴, V. Kao Kotamarthi⁵, Michael L. Stein¹ and Elisabeth J. Moyer¹
¹University of Chicago
²University of Cincinnati
³University of California at Berkeley
⁴Argonne National Lab

Future climate projections robustly suggest increased precipitation in most areas on the globe, but interannual variations in precipitation are generally as large as projected changes. In a study of the contiguous United States, we investigate how these interannual
variations relate to projected mean changes, in terms of the characteristics of individual precipitation events. We analyze high spatio-temporal resolution precipitation from reanalysis, from reanalysis-driven dynamically downscaled model simulations, and from similar simulations driven by GCM output in present and future climate conditions. In each case we apply a rainstorm identification and tracking algorithm (Chang et al. 2016) to decompose the sources of changes or interannual variations in total precipitation into four characteristics: intensity, size, duration and number. Long-term increases in mean precipitation in the study area are similar in summer and winter and are driven largely by increased precipitation intensity, modulated by reduced rainstorm size (Chang et al 2016). (Winter rainstorms are also reduced in number but increased in duration.) We find here that present-day interannual variability is associated with very different variations in rainstorm characteristics. Anomalously wet present-day winters involve precipitation events of increased intensity, size, and number but decreased duration. We link these interannual variations with related large-scale physical processes (including moisture fluxes and large-scale atmospheric circulation), and discuss how these may change in a warming climate.

Statistics-Based Compression of Global Wind Fields

Jaehong Jeong\(^1\), Stefano Castruccio\(^2\), Paola Crippa\(^2\) and Marc Genton\(^2\)

\(^1\)King Abdullah University of Science and Technology
\(^2\)Newcastle University

jaehong.jeong@kaust.edu.sa

Wind has the potential to make a significant contribution to future energy resources; however, the task of locating the sources of this renewable energy on a global scale with climate models, along with the associated uncertainty, is hampered by the storage challenges associated with the extremely large amounts of computer output. Various data compression techniques can be used to mitigate this problem, but traditional algorithms deliver relatively small compression rates by focusing on individual simulations. We propose a statistical model that aims at reproducing the data-generating mechanism of an ensemble of runs by providing a stochastic approximation of global annual wind data and compressing all the scientific information in the estimated statistical parameters. We introduce an evolutionary spectrum approach with spatially varying parameters based on large-scale geographical descriptors such as altitude to better account for different regimes across the Earth’s orography. We consider a multi-step conditional likelihood approach to estimate the parameters that explicitly account for nonstationary features while also balancing memory storage and distributed computation, and we apply the proposed model to more than 18 million points on yearly global wind speed. The proposed model achieves compression rates that are orders of magnitude higher than those achieved by traditional algorithms on yearly-averaged variables, and once the statistical model is fitted, decompressed runs can be almost instantaneously generated to better assess wind speed uncertainty due to internal variability.

Approaches for Massive Spatial Data and Applications in Remote Sensing

Emily Kang and Pulong Ma

University of Cincinnati

kangel@ucmail.uc.edu

With the development of new remote sensing technology, large or even massive spatial datasets from Earth observation become available. Statistical analysis of such data is challenging. We propose a semiparametric approach to modeling and inference for massive spatial datasets. In particular, a Gaussian process with additive components is considered, with its covariance structure coming from two components: one part is flexible without assuming a specific parametric covariance function but is able to achieve dimension reduction; the second part is parametric and simultaneously induces sparsity. The inference algorithm for parameter estimation and spatial prediction is devised. The method is applied to simulated data and a massive dataset of sea surface temperature acquired from NASA’s Terra satellite. The results demonstrate the computational and inferential benefits of the proposed method over competing methods and show that our method is more flexible and more robust against model misspecification. Other applications and extensions will also be discussed.

Modeling Precipitation Extremes using Log-Histospline

\(*\)Whitney Huang\(^1\), Doug Nychka\(^2\) and Hao Zhang\(^2\)

\(^1\)Purdue University
\(^2\)National Center for Atmospheric Research

huang251@purdue.edu

One of the commonly used approaches to modeling univariate extremes is the peaks-over-threshold (POT) method. The POT method models exceedances over a (sufficiently high/low) threshold as a generalized Pareto distribution (GPD). To apply this method, a threshold has to be chosen and the estimates might be sensitive to the chosen threshold. Here we propose an alternative, the "Log-Histospline", to explore modeling the tail behavior and the remainder of the density in one step using the full range of the data. Log-Histospline applies smoothing spline on a finely binned histogram of the log transformed data to estimate its log density. By construction, we are able to preserve the polynomial tail behavior, a feature commonly observed in daily rainfall data. The Log-Histospline can be extended to the spatial setting by treating the marginal (log) density at each location as spatially indexed functional data, and perform a dimension reduction and spatial smoothing. We illustrate the proposed method by analyzing precipitation data from both regional climate model output (North American Regional Climate Change and Assessment Program (NARCCAP)) and weather stations in China.

Session 3: Recent advances in statistical methods for brain connectome networks

Estimating Brain Pathway Effects Using Large-scale Multilevel Models

Xi Luo and Yi Zhao

Brown University

xi.rossi.luo@gmail.com

The brain can be conceptualized as a dynamic network of connected nodes, and information, such as external stimuli, is processed while passing through series of brain nodes that form pathways. Structural equation modeling (SEM) is usually employed to quantify the causal information flow. However, strong empirical evidences suggest that fMRI measured activities are influenced by stimulus unrelated factors, and this violates the causal assumption of no unmeasured confounding. We propose a new multi-layer SEM framework that provides valid causal inference even if structured unmeasured confounding factors are present. We develop a constrained optimization approach to compute the model coefficients, and analyze the nonidentifiability issue due to unmeasured confounding. Because this model leverages large and complex data structures to
help remove the confounding effects, without performing sensitivity analysis, we prove that the causal effects are identifiable and our estimator is asymptotically consistent with the parametric rate. Using extensive simulated data and a large cohort fMRI dataset, we demonstrate the improvement of our approach over existing methods.

Robust Graph Change-point Detection for Brain Evolvement Study
*Fang Han*¹, Xi Chen², Honglang Wang³, Lexin Li³ and Brian Caffo⁴

¹University of Washington
²New York University
³Indiana University-Purdue University Indianapolis
⁴University of California, Berkeley

This article investigates brain structural evolvement from resting-state functional magnetic resonance imaging. The brain structure is characterized by a series of Gaussian graphical models. The graphs correspond to different subjects, are aligned by, e.g., the ages of the subjects, and need to be estimated from the subject level data. We propose a robust data-driven method for inferring the structural changes of these graphs through a three-step procedure. First, we employ a kernel-smoothing approach to estimate multiple graphs at different ages simultaneously. Secondly, we summarize graphical information, such as the number of edges, global and local efficiency, for each estimated graph, and align them as a curve. Lastly, we propose a robust least-absolute-deviation (LAD) type penalization procedure with the fused Lasso (FL) penalty, named LAD-FL, to infer the change-points in those graph summary metrics. Our method is theoretically well understood, and results show that it could effectively capture the brain evolvement pattern.

A Network-Object Method to Uncover Hidden Disease-Related Brain Connectome
Shuo Chen
University of Maryland, College Park

This article investigates brain structural evolvement from resting-state functional magnetic resonance imaging. The brain structure is characterized by a series of Gaussian graphical models. The graphs correspond to different subjects, are aligned by, e.g., the ages of the subjects, and need to be estimated from the subject level data. We propose a robust data-driven method for inferring the structural changes of these graphs through a three-step procedure. First, we employ a kernel-smoothing approach to estimate multiple graphs at different ages simultaneously. Secondly, we summarize graphical information, such as the number of edges, global and local efficiency, for each estimated graph, and align them as a curve. Lastly, we propose a robust least-absolute-deviation (LAD) type penalization procedure with the fused Lasso (FL) penalty, named LAD-FL, to infer the change-points in those graph summary metrics. Our method is theoretically well understood, and results show that it could effectively capture the brain evolvement pattern.

Detecting copy number variations from next-generation sequencing data via a Bayesian procedure
*Guan-Hua Huang*¹ and Yu-Chung Wei²

¹National Chiao Tung University
²Feng Chia University

Copy number variations (CNVs) are genomic structural mutations with abnormal gene fragment copies. Read depth signal mirrors the variants directly from the next generation sequencing. Some tools have been published to predict CNVs by depths, but most of them just apply to a specific data type. Providing a comprehensive CNV detection algorithm that can easily make use of a variety of data types is difficult but valuable. In this paper, we will develop a Copy Number variation detection tool by a Bayesian procedure, CONY, which adopts a hierarchical model and an efficient reversible jump Markov chain Monte Carlo inference algorithm for whole genome sequencing read depths data. CONY can be applied not only to an individual for estimating the absolute number of copies but also to case-control pairs for detecting patient specific variations. Real data from the 1000 Genomes Project will be analyzed. We will also evaluate the performance of CONY and compare it with competing approaches via simulations.

Estimating links of a network from time to event data
*Tso-Jung Yen*¹, Zong-Rong Lee², Yi-Hau Chen¹, Yu-Min Yen³ and Jing-Shiang Hwang⁴

¹Institute of Statistical Science, Academia Sinica
²Institute of Sociology, Academia Sinica
³National Chengchi University
⁴Institute of Sociology, Academia Sinica

In this paper we develop a statistical method for identifying links of a network from time to event data. This method models the hazard function of a node conditional on event time of other nodes, parameterizing the conditional hazard function with the links of the network. It then estimates the hazard function by maximizing a pseudo partial likelihood function with parameters subject to a user-specified penalty function and additional constraints. To make such estimation robust, it adopts a pre-specified risk control on the number of false discovered links by using the Stability Selection method. Simulation study shows that under this hybrid procedure, the number of false discovered links is tightly controlled while the true links are well recovered. We apply our method to estimate a political cohesion network that drives donation behavior of 146 firms from the data collected during the 2008 Taiwanese legislative election. The results show that firms affiliated with elite organizations or firms of monopoly are more likely to diffuse donation behavior. In contrast, firms belonging to technology industry are more likely to act independently on donation.

Bayesian Empirical Likelihood Based Analysis for Patterned Missing data.
Sanjay Chaudhuri
Department of Statistics and Applied Probability, National University of Singapore

In recent times empirical likelihood based method has been frequently used in the analysis of patterned missing data. Such data may arise from two-phase designs, observations with surrogate variables or in the problems of data augmentation, where the information available in a sample is augmented with same variables obtained from a larger database. In these cases, the class of fully parametric models are relatively small and often quite difficult to specify. In frequentist analysis, an alternative empirical likelihood based formulation has been popular. We discuss an empirical likelihood based Bayesian analysis for such datasets.
Aggregating Information from Unlabeled Heterogenous Sources
Haoyang Liu and Chao Gao
University of Chicago
chaogao@galton.uchicago.edu

We consider the problem of estimating an unknown parameter with observations sampled from heterogeneous sources. Suppose one has the knowledge of the quality of each source, then an optimal procedure is maximizing weighted likelihood functions. In many real application settings, one does not know the quality of the data source, nor does one know which source generates which data observation. To accommodate that situation, we propose an algorithm to aggregate information in an adaptive way. The optimality and some variants of the algorithm will be discussed in the settings of PCA, linear regression and other models.

High Dimensional Propensity Score via Covariate Balancing
Yang Ning, Sida Peng and Kosuke Imai

This paper studies how to infer the average treatment effect (ATE) in high dimensional observational studies, where the number of pre-treatment covariates is much larger than the sample size. To account for the treatment assignment, we exploit the classical Horvitz-Thompson estimator. The key idea is that, the high dimensional propensity score model is carefully estimated to attain a weak covariate balancing property, which relaxes the existing balancing property, which relaxes the existing balancing method in the fixed dimensional case. In addition, we extend the covariate balancing method to a more general setting where the outcome models follow the generalized linear models. We show that, in high dimensions, the resulting estimator of ATE is sample bounded, and consistent, asymptotically normal and semiparametrically efficient. The asymptotic results are fully supported by the results of empirical studies, which include extensive simulation studies and real data analysis that compare the performance of the proposed estimator with other existing methods.

A Joint Optimal Design for Functional Data with Application to Scheduling Ultrasound Scans
So Young Park, Luo Xiao, Jayson Wilbur, Ana-Maria Staicu and N.L. Shasha

We study a joint optimal design problem for sampling functional data. The goal is to find optimal time points for sampling functional data so that the underlying truth function as well as a scalar outcome can be accurately predicted. The problem is motivated by a fetal growth study, where the ob-jective is to determine the optimal times to collect ultrasound measurements and the number of ultrasound measurements that are needed to recover fetal growth trajectories and to predict child birth outcomes. Under the frameworks of functional principal component analysis and functional linear models, we formulate the joint design problem as an optimization criterion. We implement the proposed joint design via a pilot study and also propose a penalized method for selecting the number of optimal sampling points. Performance of the proposed method is thoroughly investigated via a simulation study and by its application to the fetal ultrasound.

An Iterative Penalized Least Squares Approach to Sparse Canonical Correlation Analysis
Xin Zhang
Florida State University
henry@stat.fsu.edu

Canonical correlation analysis (CCA) is a classical tool that explores the dependency of two multivariate random variables and extracts canonical pairs of linear combinations that are highly correlated. Therefore, many recent studies aim to generalize the classical CCA to high-dimensional settings. We propose a new sparse CCA (SCCA) method that recasts high-dimensional CCA as an iterative penalized least squares problem. Thanks to the new penalized least squares formulation, our SCCA method directly penalizes and estimates the sparse CCA directions with efficient algorithms.

Critical Steps for Composite Endpoint Analysis
Jerry Li
Merck & Co., Inc.
jing.li20@merck.com

Composite endpoints have been widely used in clinical trials. A composite endpoint consists of multiple component endpoints that are combined for the investigational assessment of drug treatment effect. Per ICH E9 guideline the selection of composite endpoints has been considered as an important strategy for registration of pharmaceuticals. When designed and analyzed appropriately, composite endpoints can have many advantages such as achieving study goals with smaller and less costly trials and improved statistical efficiency. Critical steps of missing data handling and endpoints derivations will be discussed in this presentation to ensure accurate and efficient analyses of composite endpoints.

Win ratio and its generalized analytic solution for analyzing a composite endpoint considering the clinical importance order among components
Gashong Dong, Di Li, Steffen Ballerstedt, Marc Vandemeulebroecke and Duolao Wang
1iStats Inc., USA

Critical steps of missing data handling and endpoints derivations will be discussed in this presentation to ensure accurate and efficient analyses of composite endpoints.
A composite endpoint consists of multiple components combined in one outcome. It is frequently used as the primary endpoint in cardiovascular, oncology, transplant, and other therapeutic areas. There are two main disadvantages associated with the use of composite endpoints: (a) conventional approaches treat its components equally; and (b) in time-to-event analyses, the first event occurred may not be the most important component. Pocock et al. (2012) introduced the win ratio method to address these disadvantages, among other methods developed by other researchers. The win ratio is a straightforward method that takes into account the order of importance of the different components: it compares each subject in the Treatment group with every subject in the Control group to determine who is the “winner” or the “loser” based on the prioritized components, and then it takes the ratio of the number of winners in the Treatment group to that in the Control group. During the past few years, the win ratio has been applied in the design and analysis of some clinical trials, and there are also some new methodological developments such as Luo et al. (2015), Bebu and Lachin (2016), Wang and Pocock (2016) and Dong et al. (2016). This presentation will briefly review the win ratio and its methodological developments and applications in some clinical trials; and then focus on the most recent method published by Dong et al. (2016). This method provides a generalized analytic solution that is valid for any way of defining winners, losers and ties.

Session 7: New methods development in Genetic and genomic data analysis

IMAC: A Statistical Framework for Integrating Multiple Annotations to Characterize Functional Roles

Can Yang
Hong Kong Baptist University
eeyangc@gmail.com

Recent international projects, such as the Encyclopedia of DNA Elements (ENCODE) project, the Roadmap project and the Genotype-Tissue Expression (GTEx) project, have generated vast amounts of genomic annotation data, e.g., epigenome and transcriptome. There is great demand for effective statistical approaches to integrate genomic annotations with the results from genome-wide association studies. In this talk, we introduce a statistical framework, named IMAC, for integrating multiple annotations to characterize functional roles of genetic variants that underlie human complex phenotypes. For a given phenotype, IMAC can adaptively incorporates relevant annotations for prioritizing of genetic risk variants, allowing nonlinear effects among these annotations, such as interaction effects between genomic features. Specifically, we assume that the prior probability of a variant associated with the phenotype is a function of its annotations F(X), where X is the collection of the annotation status and F(X) is an ensemble of decision trees, i.e., 

\[ F(X) = \sum f_k(X) \] and \( f_k(X) \) is a shallow decision tree. We have developed an efficient EM-Boosting algorithm for model fitting, where a shallow decision tree grows in a gradient-Boosting manner (Friedman J. 2001) at each EM-iteration. Our framework inherits the nice property of gradient boosted trees: (1) The gradient accent property of the Boosting algorithm naturally guarantees the convergence of our EM-Boosting algorithm. (2) Based on the fitted ensemble \( \hat{F}(X) \), we are able to rank the importance of annotations, measure the interaction among annotations and visualize the model via partial plots (Friedman J. 2005). Using IMAC, we performed integrative analysis of genome-wide association studies on human complex phenotypes and genome-wide annotation resources, e.g., Roadmap epigenome. The analysis results revealed interesting regulatory patterns of risk variants. These findings deepen our understanding of genetic architectures of complex phenotypes. The statistical framework developed here is also broadly applicable to many other areas for integrative analysis of rich data sets.

Robust Genetic Prediction of Complex Traits with the Latent Dirichlet Process Regression Models

Ping Zeng and Xiang Zhou
University of Michigan
xzhou@umich.edu

There has been a growing interest in using genotype data to perform genetic prediction of complex traits. Accurate genetic prediction can facilitate genomic selection in animal and plant breeding programs, and can aid in the development of personalized medicine in humans. Because most complex traits have a polygenic architecture and are each influenced by many genetic variants with small effects, accurate genetic prediction requires the development of polygenic methods that can model all genetic variants jointly. Many recently developed polygenic methods make parametric modeling assumptions on the effect size distribution and different polygenic methods differ in such effect size assumption. However, depending on how well the effect size distribution assumption matches the unknown truth, existing polygenic methods can perform well for some traits but poorly for others. To enable robust phenotype prediction performance across a range of phenotypes, we develop a novel polygenic model with a flexible assumption on the effect size distribution. We refer to our model as the latent Dirichlet Process Regression (DPR). DPR relies on the Dirichlet process to assign a prior on the effect size distribution itself, is non-parametric in nature, and is capable of inferring the effect size distribution from the data at hand. Because of the flexible modeling assumption, DPR is able to adapt to a broad spectrum of genetic architectures and achieves robust predictive performance for a variety of complex traits. We compare the predictive performance of DPR with several commonly used polygenic methods in simulations. We further illustrate the benefits of DPR by applying it to predict gene expressions using cis-SNPps, to conduct PrediXcan based gene set test, to perform genomic selection of four traits in two species, and to predict five complex traits in a human cohort. Our method is implemented in the DPR software, freely available at www.xzlab.org/software.html.

SynthEx: A synthetic-normal based DNA sequencing tool for copy number alteration detection

Mengjie Chen, Grace O. Silva, Marni B. Siegel, Lisle E. Mose, Joel S. Parker, Wei Sun and Charles M. Perou

1the University of Chicago
2University of North Carolina - Chapel Hill
3Fred Hutchinson Cancer Research Center
mengjiechen@uchicago.edu

Changes in the quantity of genetic material, known as somatic copy number alterations (CNAs), can drive tumorigenesis. Many methods exist for assessing CNAs using microarrays, but considerable technical issues limit current CNA calling based upon DNA sequencing. We present SynthEx, a novel tool for detecting CNAs from whole exome and genome sequencing. SynthEx utilizes a "synthetic-normal" strategy to overcome technical and financial issues. In terms of accuracy and precision, SynthEx is highly com-
parable to array-based methods and outperforms sequencing-based CNA detection tools. SynthEx robustly identifies CNAs using sequencing data without the additional costs associated with matched normal specimens.


Jin Liu
DUKE-NUS Medical School [jin.liu@duke-nus.edu.sg]

Most existing methods for GWAS data analysis require individual-level genotype data as their input. However, it is often not easy to get access to individual-level data, due to many practical issues, such as privacy protection and disagreement on data-sharing among multiple research groups. In this talk, we introduce SSREM, a Summary-Statistics-based approach to estimating heritability, co-heritability and effect sizes in GWAS data analysis. This is achieved by Bayesian analysis with the standard random-effect prior and a summary-statistics-based likelihood function. We have implemented a parallel Gibbs sampling strategy, which allows us to handle genome-wide-scale datasets. Our analysis results suggest that summary-statistics-based analysis can achieve comparable performance to individual-level data analysis.

Session 8: Recent developments in multiple testing

Copula-based multiple hypothesis testing
Thorsten Dickhaus
University of Bremen, Institute for Statistics [dickhaus@uni-bremen.de]

We are concerned with simultaneous testing of a family of null hypotheses under a single statistical model. In this, we assume that the individual tests are carried out by means of (marginal) p-values and that these p-values, regarded as random variables, are dependent. We model the dependency structure among the p-values by an unknown copula function, which we regard as an infinite-dimensional nuisance parameter of the joint model for the vector of p-values. Pre-estimation of the p-value copula leads to empirically calibrated multivariate multiple tests for control of the family-wise error rate or the false discovery rate, respectively. Finally, we deal with estimating the proportion of true null hypotheses under copula dependence by making use of a marginal parametric bootstrap approach.

False Discoveries Occur Early on the Lasso Path
\textsuperscript{\dag}Weijie Su\textsuperscript{1}, Malgorzata Bogdan\textsuperscript{2} and Emmanuel Candès\textsuperscript{3}
\textsuperscript{1}University of Pennsylvania
\textsuperscript{2}University of Wroclaw,
\textsuperscript{3}Stanford University [suw@wharton.upenn.edu]

In regression settings where explanatory variables have very low correlations and there are relatively few effects, each of large magnitude, we expect the Lasso to find the important variables with few errors, if any. In this talk we show that in a regime of linear sparsity—meaning that the fraction of variables with a non-vanishing effect tends to a constant, however small—this cannot really be the case, even when the design variables are stochastically independent. We demonstrate that true features and null features are always interspersed on the Lasso path, and that this phenomenon occurs no matter how strong the effect sizes are. We derive a sharp asymptotic trade-off between false and true positive rates or, equivalently, between measures of type I and type II errors along the Lasso path. This trade-off states that if we ever want to achieve a type II error (false negative rate) under a critical value, then anywhere on the Lasso path the type I error (false positive rate) will need to exceed a given threshold so that we can never have both errors at a low level at the same time. Our analysis uses tools from approximate message passing (AMP) theory as well as novel elements to deal with a possibly adaptive selection of the Lasso regularization parameter.

Robust Estimation: New Perspective, Theory and Applications to Large-Scale Multiple Testing
\textsuperscript{\dag}Wenxin Zhou, Koushiki Bose, Jianqing Fan and Han Liu
Princeton University [stevewenxin@hotmail.com]

Heavy-tailed errors impair the accuracy of the least square estimate, which can be spoiled by a single grossly outlying observation. As argued in the seminar work of Peter Huber in 1973 [Ann. Statist. 1 (1973) 799(821)], robust alternatives to the method of least squares are sorely needed. To achieve robustness against heavy tailedness, we revisit the Huber estimator from a new perspective by letting the regularization parameter involved diverge with the sample size. In this paper we develop nonasymptotic concentration results for such an adaptive Huber estimator, namely, the Huber estimator with the regularization parameter adapted to sample size, dimension and the variance of the noise. Specifically, we obtain a sub-Gaussian-type deviation inequality and a nonasymptotic Bahadur representation in the presence of heavy-tailed errors. The nonasymptotic results further yield two important normal approximation results, the Berry-Esseen inequality and Cramer-type moderate deviation. As an important application to large-scale simultaneous inference, we apply these robust normal approximation results to analyze a dependence-adjusted multiple testing procedure for moderately heavy-tailed data. It is shown that the robust, dependence-adjusted procedure asymptotically controls the overall false discovery proportion at the nominal level under mild moment conditions. Thorough numerical results on both simulated and real datasets are also provided to back up our theory.

Optimal Rates and Tradeoffs in Multiple Testing
Maxim Rabinovich, \textsuperscript{\dag}Aaditya Ramdas, Michael Jordan and Martin Wainwright
UC Berkeley [arandas@berkeley.edu]

Multiple hypothesis testing is a central topic in statistics, but despite abundant work on the false discovery rate (FDR) and the corresponding Type-II error concept known as the false non-discovery rate (FNR), a fine-grained understanding of the fundamental limits of multiple testing has not been developed. Our main contribution is to derive a precise non-asymptotic tradeoff between FNR and FDR for a variant of the generalized Gaussian sequence model. Our analysis is flexible enough to permit analyses of settings where the problem parameters vary with the number of hypotheses n, including various sparse and dense regimes (with o(n) and O(n) signals). Moreover, we prove that the Benjamini-Hochberg algorithm as well as the Barber-Candes algorithm are both rate-optimal up to constants across these regimes.

Session 9: Novel Analysis of Big Data arising from mHealth and Sensor Technology

SMART Design Considerations for mHealth Interventions
Bibhas Chakraborty
A sequential multiple-assignment randomized trial (SMART) is a novel trial design custom-made to evaluate treatment sequences or strategies, rather than stand-alone treatments. Many mobile health (mHealth) interventions are time-varying, adaptive interventions, and as such, SMART designs offer great promise in this context. In this talk, we will discuss the common considerations relating to using a SMART design for mHealth. In particular, through a simulation study, we will demonstrate the merit of a SMART, compared to a standard randomized controlled trial (RCT), for mHealth interventions. In addition, we will consider non-inferiority and equivalence testing within a SMART design framework, which should facilitate the uptake of SMART in the mHealth setting.

Transforming N-of-1 m-Health Data into Actionable Insights: A Case Study of Stress-Behavior Pathway

Pei-Yan Hsueh1, Xinyu Hua2, Min Qian2 and Ying Kuen Cheung3
1IBM TJ Watson Research Center
2Columbia University
3Pei-Yun Hsueh@gmail.com

In the domain of healthcare, the rise of consumer awareness and the prevalence of wearables and sensor technologies have enabled the accumulation of m-health data and offered promise to precision health applications. Making individual recommendations based on the accumulated m-health data, however, incurs multi-level challenges. These include applying population-level evidence at an individual level and inducing actionable feedback from sequential observations at each potential touchpoint in the intermediate stages. The goal of this talk is to introduce the development of a statistical learning framework that can help capture individual predictive pathways from the observed behaviors and proxy outcomes. For example, we leverage the observations of past exercise-stress behavioral pathways to induce policies for physical activity recommendations that can reduce a user’s perceived stress over a given time horizon. Ideographic N-of-1 models are compared with the traditional nomothetic (one-size-fits-all) model. Moreover, a new form of reinforcement learning method, multi-stage threshold Q-learning (mTQL) is introduced to incorporate threshold selection into the learning process. Results from both the observational and Monte Carlo simulation studies indicate that the mTQL-based learning framework brings out actionable insights from m-health data and can empower patients in precision behavioral intervention and cognitive care management applications.

Flexible Functional Clustering of Accelerometer Data via Transformed Input Variables

Hee-Seok Oh1, Yaeji Lim2 and Ying Kuen Cheung3
1Seoul National University
2Pukyong National University
3Columbia University
heeseok.oh@gmail.com

This paper considers the clustering problem of physical activity data measured by computerized accelerometer. Classical methods such as K-means and partitioning around medoids are not efficient in handling the accelerometer data that are high-dimensional and have inherent multiscale structures. Existing functional clustering approaches do not naturally utilize the dynamic structures of the accelerometer data that may be necessary to form homogenous clusters in a meaningful way. This paper introduces new input variables for clustering the accelerometer data based on rank-based transform and thick-pen transform, which reflect specific structures of the data such as the amount and the pattern of physical activity while preserving a functional form. The proposed clustering methods are obtained by coupling the transformed input variables with functional clustering that considers a marginal representation of the data for building clustering criteria. We suggest several clustering schemes using the proposed methods and apply the schemes to a real data set of 365 subjects. Simulation study is performed to evaluate the empirical performance of the proposed methods, which are shown to be superior to some existing methods.

Center for Digital Health at UMass: Research and Opportunities

Jong Soo Lee
University of Massachusetts Lowell
jongsoo.lee@umass.edu

The Center for Digital Health (CHD) at the University of Massachusetts was established recently to foster collaboration across the University of Massachusetts campuses (including the Medical School campus) and engage in cutting-edge medical research. The center will focus on modern digital health issues including mHealth and big data. Specifically, the three main research objectives are (1) designing scalable pervasive healthcare monitoring, rehabilitation, and public health systems, (2) building high performance networking and computing infrastructure for health data transmission and computation, and (3) developing novel algorithms and systems for big data analytics in healthcare (with emphasis on deep learning). I will talk about some of the current and recent researches related to the center and discuss some of the opportunities for statisticians going forward.

Session 10: Recent advances in methods for EHR based research

Constructing stabilized dynamic treatment regimes using electronic health record data

Yingyi Zhao
Fred Hutchinson Cancer Research Center
yqzhao@frehutch.org

Dynamic treatment regimes (DTRs) are sequential decision regimes for individual patients that can adapt over time to an evolving illness. The goal is to find the DTRs tailored to individual characteristics that lead to the best long term outcome if implemented. We introduce a general learning framework on stabilized dynamic treatment regimes (SDTRs), where we can make stabilized decisions over time using the repeated measured information on the same variable. The proposed method is based on directly maximizing an estimator of the expected long-term outcome over all constrained SDTRs. Experimental studies showed a superior performance of the proposed method.

Constructing Dynamic Treatment Regimes in Infinite-Horizon Settings

Ashkan Ertefaie
University of Rochester
ertefaie

Existing methods for estimating optimal dynamic treatment regimes is limited to cases where investigators are interested in optimizing a utility function over a fixed period of time, i.e., finite-horizon. In this manuscript, we develop an inferential procedure based on temporal difference residuals for optimal dynamic treatment regimes in infinite-horizon settings, where there is no a priori fixed end of follow-up point. The proposed method can be used to estimate the
Predicting complex phenotypes in electronic health records with semiparametric CCA

Denis Agniel and Tianxi Cai
1 RAND Corporation
2 Harvard University
dagniel@rand.org

Studying multiple outcomes simultaneously allows researchers to begin to identify underlying factors that affect all of a set of diseases (i.e., shared etiology) and what may give rise to differences in disorders between patients (i.e., disease subtypes). In this work, our goal is to build risk scores that are predictive of multiple phenotypes simultaneously and identify subpopulations at high risk of multiple phenotypes. Such analyses could yield insight into etiology or point to treatment and prevention strategies. The standard canonical correlation analysis (CCA) can be used to relate multiple continuous outcomes to multiple predictors. However, especially in electronic health records research and other settings where data are collected passively, phenotypes may include a diverse range of data types, including binary, continuous, ordinal, and censored variables. When phenotypes are diverse in this way, standard CCA is not possible and no methods currently exist to model them jointly. In the presence of such complications, we propose a semiparametric CCA method to develop risk scores that are predictive of multiple phenotypes. To guard against potential model mis-specification, we also propose a nonparametric calibration method to identify subgroups that are at high risk of multiple disorders. Our method opens the door to synthesizing a wide array of EHR data sources for the purposes of joint prediction.

Bias reduction methods for EHR data based association studies

Yong Chen
University of Pennsylvania
ychen123@upenn.edu

In this talk, we present a novel Prior knowledge guided Integrated likelihood Estimation method (PIE) for bias correction in Electronic Health Records (EHR)-based association studies, and evaluates the performance of the proposed method by comparing it to two methods that are commonly used in current practice. We conducted simulation studies and data analysis of real EHR-derived data on diabetes from Kaiser Permanente Washington to evaluate the strengths and limitations of the proposed method. The proposed method effectively leverages available information on phenotyping accuracy to construct a prior distribution for sensitivity and specificity, and incorporates this prior information through the integrated likelihood for bias reduction.

Session 11: Recent Advances in High-dimensional Inference

Regression coefficients clustering in high dimensional data

Jungh Ki, Hongtu Zhu and Xiao Wang
1 University of Texas MD Anderson Cancer Center
2 Purdue University
jkim30@mdanderson.org

In this study, we introduce a fast and efficient strategy for exploring the heterogeneity of coefficients of multiple studies in high-dimensional regression. We often integrate multiple datasets of similar studies in biomedical research to increase sample sizes for statistical power improvement. Linking data across datasets is challenging due to the heterogeneity in multiple studies, which results in biased estimation and misleading inference. A fusion penalty approach allows to identify homogeneous or heterogeneous parameters of regression analysis. By incorporating the linear transformations, we learn parameter heterogeneity in data fusion. We demonstrate the performance of our approach with comparison to an existing method and provide an analysis of high-dimensional gene expression data.

High-dimensional inference robust to sparsity

Jelena Bradic and Yinchu Zhu
University of California, San Diego
jbradic@ucsd.edu

In high-dimensional linear models the sparsity assumption is typically made, stating that most of the parameters are equal to zero. Under the sparsity assumption, estimation and, recently, inference have been well studied. However, in practice, sparsity assumption is not checkable and more importantly is often violated, with a large number of covariates expected to be associated with the response, indicating that possibly all, rather than just a few, parameters are non-zero. A natural example is a genome-wide gene expression profiling, where all genes are believed to affect a common disease marker. We show that existing inferential methods are sensitive to the sparsity assumption, and may, in turn, result in the severe lack of control of Type-I error. In this article, we propose a new inferential method, named CorrT, which is robust and adaptive to the sparsity assumption. CorrT is shown to have Type I error approaching the nominal level and Type II error approaching zero, regardless of how sparse or dense the model. In fact, CorrT is also shown to be optimal whenever sparsity holds. Numerical and real data experiments show a favorable performance of the CorrT test compared to the state-of-the-art methods.

Homogeneity Test of Covariance Matrices with High-Dimensional Longitudinal Data

Ping-Shou Zhong and Runze Li
1 Michigan State University
2 Penn State University
pszhong@stt.msu.edu

High-dimensional longitudinal data such as time-course microarray data are now widely available. One important feature of such data is that, for each individual, high-dimensional measurements are repeatedly collected over time. Moreover, these measurements are spatially and temporally dependent which, respectively, refers to dependence within each particular time point and among different time points. This paper focuses on testing the homogeneity of covariance matrices of high-dimensional measurements over time against the change-point type alternatives. We allow the dimension of measurements (p) to be much larger than the number of individuals (n). Specifically, a test statistic for the equivalence of covariance matrices is proposed and the asymptotic normality is established. In addition to testing, an estimator for the location of the change point is given whose rate of convergence is established and shown to depend on p, n and the signal-to-noise ratio. The proposed method is extended to locate multiple change points by applying a binary segmentation approach, which is shown to be consistent under some mild conditions. The proposed testing procedure and change-point

Abstracts
identification methods are able to accommodate both spatial and temporal dependence. Simulation studies and an application to a time-course microarray data set are presented to demonstrate the performance of the proposed method.

A New Scope of Penalized Empirical Likelihood with High-Dimensional Estimating Equations

♦ Jinyuan Chang\(^1\), Cheng Yong Tang\(^2\) and Tong Tong Wu\(^3\)

\(^1\)Southwestern University of Finance and Economics
\(^2\)Temple University
\(^3\)University of Rochester

jinyuan.chang@unimelb.edu.au

Statistical methods with empirical likelihood (EL) are appealing and effective especially in conjunction with estimating equations through which useful data information can be adaptively and flexibly incorporated. It is also known in the literature that EL approaches encounter difficulties when dealing with problems having high-dimensional model parameters and estimating equations. To overcome the challenges, we begin our study with a careful investigation on high-dimensional EL from a new scope targeting at estimating a high-dimensional sparse model parameters. We show that the new scope provides an opportunity for relaxing the stringent requirement on the dimensionality of the model parameter. Motivated by the new scope, we then propose a new penalized EL by applying two penalty functions respectively regularizing the model parameters and the associated Lagrange multipliers in the optimizations of EL. By penalizing the Lagrange multiplier to encourage its sparsity, we show that drastic dimension reduction in the number of estimating equations can be effectively achieved without compromising the validity and consistency of the resulting estimators. Most attractively, such a reduction in dimensionality of estimating equations is actually equivalent to a selection among those high-dimensional estimating equations, resulting in a highly parsimonious and effective device for high-dimensional sparse model parameters. Allowing both the dimensionalities of model parameters and estimating equations growing exponentially with the sample size, our theory demonstrates that the estimator from our new penalized EL is sparse and consistent with asymptotically normally distributed nonzero components. Numerical simulations and a real data analysis show that the proposed penalized EL works promisingly.

Session 12: Survival Analysis and Genetics

Selection models for enhancing power of tests of genetic associations in family studies

Richard Cook
University of Waterloo

rjcook@uwaterloo.ca

Studies about the genetic basis for disease are often conducted by selecting affected individuals called probands, recruiting and examining family members, and assessing the within-family association in one or more features of the disease process (e.g., presence, age of onset, severity). Models which accommodate within-family dependencies offer a valid and potentially powerful way of assessing the effect of genetic markers on disease occurrence provided the biased sampling scheme is suitably accounted for. We develop selection models which use information available on the probands to select families in a more efficient way when analyses are based on correlated times of disease onset. Tests based on the marginal onset time distribution and second-order dependence parameters are developed within the frameworks of composite likelihood and estimating functions. An application to a motivating arthritis family study is given for illustration.

Assessment of familial genetic cancer risk in the presence of competing risks

Lakhal-Chaieb MHamed Lajmi
Laval University

lakhal@mat.ulaval.ca

In this presentation, an association model to estimate the penetrance (risk) of multiple cancers in the presence of competing risks is proposed. The association between multiple events is modelled via a pair of opulas. This work is motivated by the analysis of breast and ovarian cancers for women with a mutation in RCA1/2 genes. The proposed inference procedure is adapted to handle selection bias, induced by the data collection protocol of the data at hand. The proposed methods are illustrated with data from the consortium of Investigators of Modifiers of BRCA1/2 (CIMBA).

An additive hazards model for gene level association analysis of survival traits of complex disorder

♦ Chi-yan Chiu\(^1\), Richard Cook\(^2\) and Ruzong Fan\(^3\)

\(^1\)National Institutes of Health, \(^2\)University of Waterloo, \(^3\)Georgetown University Medical Center

chiuchiyang@gmail.com

Based on counting processes and Doob-Meyer decomposition of submartingales and functional regression models, we propose an additive hazards model for gene level association analysis of survival traits of complex disorders. The additive hazards model can overcome proportional hazards assumptions of Cox models and is more flexible to model association with time/age. Association between genetic markers and eye disease will be investigated.

Pedigree-based Association Analysis of Survival Traits via Functional Regressions

Ruzong Fan
Georgetown University Medical Center

rf740@georgetown.edu

Functional regression models have been developed for unrelated samples to test for association between a quantitative or a dichotomous or a survival trait and genetic variants in a major gene region. In most situations, these models have higher power than well-known kernel-based methods (SKAT and SKAT-O). Here we extend this approach to accommodate family-based data using functional Cox mixed models (FCoxME). The FCoxME models the effect of major gene as fixed mean via functional data analysis techniques, the polygenic contributions as a random variation, and the correlation of pedigree members by kinship coefficients. The association between the censored trait and the major gene is tested by likelihood ratio tests (LRT). Our proposed FCoxME provides a new tool for conducting family-based research studies in public health for complex or multifactorial diseases. The models and related test statistics can be useful in the whole genome and whole exome association studies.

Session 13: Design and Analysis of Dose Finding Clinical Trials for Combination Therapy

Evaluation of false positive rate based on E-R analyses for two compounds in fixed-dose product

♦ Yuan Wang and Hao Zhu

U.S. FDA

Hao.Zhu@fda.hhs.gov
We explored the type I error rate (false positive rate) associated with exposure-response (ER) analyses for two compounds in a fixed-dose combination product through simulations. In the simulations, at least one compound was assumed to be inactive, whereas the active compound followed $E(\text{max})$ model at different concentration ranges. The simulated data were independently evaluated by pre-specified univariate or multivariate linear, log-linear models, and mixed linear log-linear models. The type I error rate was evaluated by comparing the total number of falsely identified significant slope estimates with the total number of models with successful convergence. We demonstrated that ER analyses results based on data from fixed-dose combination products at various dose levels should be interpreted with caution. A univariate analysis, even though is appropriate to guide dose selection, is inadequate to identify the active compound. Multivariate analyses can be applied to determine the active compound only when the underlying ER relationship for each compound (especially for the active compound) has been adequately defined or approximated. The false positive rate in determining a significant ER relationship is elevated, when the underlying ER relationship (especially for the active compound) is erroneously or inadequately defined. Without the assurance of the correct structural models, the identified significant ER relationship does not necessarily indicate that the compound associated with the significant slope estimate is pharmacologically active. Disclaimer: The views expressed in this paper are those of the authors and do not necessarily represent those of the FDA.

Dose-Finding for Immunotherapy Combinations using a Conditionally Autoregressive Model

Thomas Braun
University of Michigan Department of Biostatistics
tombraun@umich.edu

Phase I trials in oncology have advanced from studies of the safety of a single agent to studies of the simultaneous toxicity and efficacy of two agents. Furthermore, the new agents under study are often immunotherapeutic rather than chemotherapeutic agents. The nature of immunotherapy brings into question the adage from the era of chemotherapy that “more is better”, requiring deeper thought into how we model efficacy as a function of dose changes of one or two immunotherapeutic agents. The nature of immunotherapy brings into question the adage from the era of chemotherapy that “more is better”, requiring deeper thought into how we model efficacy as a function of dose changes of one or two immunotherapeutic agents. We demonstrate that several published modeling choices implicitly place strong correlation constraints among the dose combinations that lead to over-smoothing of the data. As an alternative, we propose the use of a conditional autoregressive (CAR) model, which allows us to model the correlation directly and leads to a direct control on the amount of smoothing that is used. We describe the general structure of CAR models and then present simulation results comparing the operating characteristics of a CAR model-based design with other existing designs. We then conclude with a discussion of several nuances of CAR models that require further study.

Designing early phase drug combination trials in practice

Ying Yuan
MD Anderson Cancer Center
yyuan@mdanderson.org

Drug combination therapy is an important approach to tackle hard to treat diseases such as cancer. In this talk, I will discuss the characteristics and challenges of designing early phase drug combination trials. I will show the importance of distinguishing the trials that aim to finding a single MTD and the trials that aim to find multiple MTDs, i.e., the MTD contour, and its implication on the trial design. Focusing on the practical implementation of drug combination trials, several intuitive and easy-to-implement approaches will be introduced. The related software will be described for designing real-world drug combination trials.

Modelling semi-attributable toxicity in dual-agent phase I trials

Graham Wheeler, Michael Sweeting, Adrian Mander, Shing Lee and Ken Cheung

1Columbia University
yc632@cumc.columbia.edu

In oncology, combinations of drugs are often used to improve treatment efficacy and/or reduce harmful side effects. Dual-agent phase I clinical trials assess drug safety and aim to discover a maximum tolerated dose combination via dose-escalation; cohorts of patients are given set doses of both drugs and monitored to see if toxic reactions occur. Dose-escalation decisions for subsequent cohorts are based on the number and severity of observed toxic reactions, and an escalation rule. In a combination trial, drugs may be administered concurrently or non-concurrently over a treatment cycle. For two drugs given non-concurrently with overlapping toxicities, toxicities occurring after administration of the first drug yet before administration of the second may be attributed directly to the first drug, whereas toxicities occurring after both drugs have been given some present ambiguity; toxicities may be attributable to the first drug only, the second drug only or the synergistic combination of both. We call this mixture of attributable and non-attributable toxicity semi-attributable toxicity. Most published methods assume drugs are given concurrently, which may not be reflective of trials with non-concurrent drug administration. We incorporate semi-attributable toxicity into Bayesian modelling for dual-agent phase I trials with non-concurrent drug administration and compare the operating characteristics to an approach where this detail is not considered. Simulations based on a trial for non-concurrent administration of intravesical Cabazitaxel and Cisplatin in early-stage bladder cancer patients are presented for several scenarios and show that including semi-attributable toxicity data reduces the number of patients given overly toxic combinations.

Session 14: Understanding the Microbiome Complexity –Genetics and Networks

Microbial network estimation using bias-corrected graphical lasso

*Duo Jiang, Yuan Jiang and Thomas Sharpton

Oregon State University
jiangd@stat.oregonstate.edu

With the increasing availability of microbiome 16S data, network estimation has become a useful approach to studying the interactions between microbial taxa. Network estimation on a set of variables is frequently explored using graphical models, in which the relationship between two variables is modeled via their conditional dependency given the other variables. In recent years, various methods for sparse inverse covariance estimation have been proposed to estimate graphical models in the high-dimensional setting, including graphical lasso. However, current methods do not address the compositional nature of microbiome data, where abundances of microbial taxa are not directly measured, but are reflected by the observed counts in an error-prone manner. Adding to the challenge is the fact that the sum of the counts within each sample, termed “sequencing depth”, is an experimental technicality, which carries no scientific information but can vary drastically across samples.
To address these issues, we develop a new approach to network estimation, which models the microbiome data using a multinomial log-normal distribution with the finite sequencing depth explicitly incorporated. We propose to improve the empirical covariance estimator via a computationally simple procedure that corrects the bias arising from the heterogeneity in sequencing depth. We then build our inverse covariance estimator on graphical lasso. We will show the advantage of our method in comparison to current approaches for inverse covariance estimation under a variety of simulation scenarios. We will also illustrate the use of our method in an application to human microbiome data set.

Compositional mediation model for microbiome study

*Michael Sohn and Hongze Li*
University of Pennsylvania

Motivated by advances in the causal inference on mediation and problems arising in the analysis of the microbiome study, we consider the effect of a treatment on an outcome transmitted through compositional mediators. A sparse mediation model for high-dimensional compositional data is proposed by utilizing the algebraic structure of a composition under the simplex space and a constrained linear regression model to achieve subcompositional coherence. Under this model, we develop estimation methods for direct and indirect effects of a randomly assigned treatment on the outcome. We applied the method to a gut microbiome dataset to investigate the effect of fat intake on body mass index (BMI) mediated through the gut microbiome composition.

A Testing Framework for Detecting Differential Microbial Networks

*Jing Ma*
1. University of Pennsylvania
2. Fudan University

Microorganisms such as bacteria do not exist in isolation but form complex ecologies with various symbiotic relationships. These microbial communities undergo differential changes that are dependent on environment or disease state. Differential analysis of microbial community structures is powerful for elucidating systematic changes during an adaptive response. In this talk, we propose a flexible model for learning the microbial community structure and introduce a testing framework for differential network mapping. Our method for testing differential network globally is particularly powerful against sparse alternatives. In addition, we develop a multiple testing procedure with false discovery rate control to infer the structure of the differential network. The proposed method is applied to a gut microbiome study in UK twins to detect the microbial interaction associated with age.

Session 15: Recent Developments in Neuroimaging Analysis

FMRI Preprocessing Changes the Statistical Properties of Your Data

*Daniel Rowe*
1. Marquette University
2. University of Alberta
3. University of Pennsylvania

In fMRI and fcMRI, the original (k-space) data measured by the MRI machine is in a form that is not usable and must go through an image reconstruction transformation process in order to turn it into an image. In addition, many data processing operations are performed on the data both before and after image reconstruction. The processing changes our image mean, variance, and may induce voxel correlation. Newer accelerated image acquisition and reconstruction processes that partially-sample the necessary (k-space) data can also induce long range correlations. This talk will review the measurement, reconstruction, and processing of fully-sampled and sub-sampled (k-space) data in addition to its resulting statistical properties.

Tensor approximation of partial quantile regression for neuroimaging data analysis

*Dengdeng Yu, Linglong Kong and Ivan Mizera*
University of Alberta

We consider the estimation problem in functional linear quantile regression in which the dependent variable is a scalar while the covariate is a function. The conditional quantile of response is modeled as a linear functional of the covariate. We are interested in the way of efficiently and effectively extracting the basis for estimating functional coefficients. There are two common approaches. One is to use the functional principal components. The other one is to use functional partial quantile basis which is just an extension of partial least squares into quantile regression. As a supervised method, partial quantile method is superior to unsupervised PCA method under certain circumstances. In this talk, we propose to use the partial quantile regression and its tensor approximation to estimate the functional effect in functional linear quantile regression model. Asymptotic properties have been studied to show the theoretical performance while both a simulation and a neuroimaging data example are investigated and some interesting findings are discovered.

Modeling dynamic brain response to pain progression using multilevel functional regression

*Haochang Shou, Victoria Arthur, Alessandra Valcarcel, John Farrar and John Detre*
University of Pennsylvania

Brain mechanism in pain management and response to medication has been a long interesting topic yet not fully understood. Previous evidences have shown correlations between changes of regional cerebral blood flow (CBF) and pain phenotypes. Arterial spin labeled (ASL) perfusion magnetic resonance imaging (MRI) is an effective technology to visualize and quantify CBF. However, most of the previous studies used simulated pain stimulus and study the immediate brain responses within a short period of time. We are motivated by a novel study design with natural pain model (impacted 3rd molar extraction) over time and pain intensity spectrum using ASL-MRI approach, accompanied with sequences of resting state fMRI and subjective pain level assessments before and after ibuprofen. Functional regression models that combine prior network information for repeated imaging scans were used to assess the association between the time-varying brain blood flow and progressive pain intensity. Connectivity between different regions of interest (ROIs) were also estimated over time and compared between treatment and placebo groups.
Session 16: Prediction and inference for high-dimensional data and nonparametric models

Generalized Fiducial Inference for High-Dimensional Sparse Additive Models
Qi Gao1, Randy Lai2 and Thomas Lee3
1UC Davis
cmlee@ucdavis.edu
2U of Maine
3U of Illinois Urbana and Champaign
zongming@wharton.upenn.edu

Generalized fiducial inference (GFI) is a relatively new tool for conducting statistical inference, although its original form was proposed by Fisher back in the 1930s. In this talk we describe our ongoing work in applying GFI to derive confidence intervals for high-dimensional sparse additive models. This is joint work with Qi Gao and Randy Lai.

Exploration of Large Networks via Fast and Universal Latent Space Model Fitting
Zhuang Ma and Zongming Ma
University of Pennsylvania
zongming@wharton.upenn.edu

Latent space models are effective tools for statistical modeling and exploration of network data. These models can effectively model real world network characteristics such as degree heterogeneity, transitivity, homophily, etc. Due to their close connection to generalized linear models, it is also natural to incorporate covariate information in them. The current paper presents two universal fitting algorithms for networks with edge covariates: one based on nuclear norm penalization and the other based on projected gradient descent. Both algorithms are motivated by maximizing likelihood for a special class of inner-product models while working simultaneously for a wide range of different latent space models, such as distance models, which allow latent vectors to affect edge formation in flexible ways. These fitting methods, especially the one based on projected gradient descent, are fast and scalable to large networks. We obtain their rates of convergence for both inner-product models and beyond. The effectiveness of the modeling approach and fitting algorithms is demonstrated on five real world network datasets for different statistical tasks.

Individualized Multi-directional Variable Selection
Xiwei Tang and Annie Qu
University of Illinois Urbana and Champaign
xtang14@illinois.edu

In this paper, we propose an individualized variable selection approach to select different relevant variables for different individuals. In contrast to conventional model selection approaches, the key component of the new approach is to construct a separation penalty with multi-directional shrinkage including zero, which facilitates the individualized modeling to distinguish a strong signal from noise one. As a byproduct, the proposed model identifies subgroups among which the individuals share similar effects, and thus improves the estimation efficiency and the personalized prediction accuracy. Another advantage of the proposed model is that it can incorporate the within-subject correlation for longitudinal data. We provide a general theoretical foundation in the double-divergence modeling where the number of subjects and the number of repeated measurements both go to infinity, and therefore yields high-dimensional individual parameters. In addition, we present the oracle property for the proposed estimator to ensure its optimal large sample property. Simulation studies and an application to the HIV longitudinal data are illustrated to compare the new approach to existing penalization methods.

Session 17: Statistical methods to advance patient tailoring and precision medicine

Individual Treatment Effect for Better Decision Making in Drug Development and Healthcare
Jia Jia, Qi Tang, Wangang Xie and Richard Rode
yu@abbvie.com

A lack of understanding of human biology creates a hurdle for the development of precision medicines. To overcome this hurdle we need to better understand the potential synergy between a given treatment (vs. placebo or active control) and various demographic or genetic factors, disease history and severity, etc., with the goal of identifying those patients at increased chance of exhibiting meaningful treatment benefit. For this reason we proposed the VG method, which combines the idea of individual treatment effect (ITE) from the Virtual Twins method (Foster et al 2013, Stat Med) and the unbiased variable selection and cutoff value determination algorithm from the GUIDE method (Loh 2015, Stat Med). Simulation results showed the VG method to have less variable selection bias than Virtual Twins and higher statistical power than GUIDE in the presence of prognostic variables with strong treatment effects. The type I error rate and predictive performance of Virtual Twins, GUIDE and VG were also compared through the use of simulation studies and a randomized clinical trial for Alzheimer’s disease.

Adaptive threshold design for precision medicine clinical trial
Meijuan Li
Food and Drug Administration
meijuan.li@fda.hhs.gov

With the advancement in molecular biology, many predictive biomarkers are developed to identify effective targeted therapies for patients. Oftentimes, drug targeted patient subgroups are unknown and must be learned through inference using observed data. In this talk, we proposed an outcome-based adaptive randomization trial design. The Bayesian adaptive model is used to update 1) to treatments enrich the allocation of each subgroup of patients to their precision and desirable treatments 2) subgroup-treatment efficacy. An early stopping rule is implemented to suspend low-performing treatments from randomization. We use simulation studies to illustrate the utility of the proposed method.

Building a Bayesian decision-theoretic framework to design biomarker-driven studies in early phase c
Danni Yu
Eli Lilly and Company
yu-danni@lilly.com

Decision theory as a subfield of Artificial Intelligence provides a quantitative strategy to guide decision makers, based on possible outcomes of all scenarios. It enables optimal selection that balances the potential gain and loss. For clinical drug development, prior to the start of a series of trials, it is crucial to characterize and quantify all the possible results that will happen after a decision. Such a quantitative strategy can potentially help decision makers select the trials delivering relatively high response rate with high predicted probability while minimizing the cost of unnecessary studies. Furthermore, a common question in clinical studies is about the sample size for a trial. Using conventional approach (controlling type I and type II errors for a typical effect size) may lead to large sample size, which is typical for phase III confirmatory studies with known tailoring biomarker. However, this frequentist approach may not be appropriate to design multiple pilot experiments or small trials for the purpose of biomarker identification in early phase studies. Therefore, the Bayesian decision-theoretic framework will be explored.
and applied to construct a tree of probabilities. The root nodes of the tree are the possible actions (i.e., a set of pilot studies or small trials for drug development). The branch nodes are the factors affecting the outcomes (such as safety, efficacy, cost, population size, etc.). The leaves are the outcomes. This presentation will focus on the benefit as well as challenges of Bayesian decision-theoretic approach in biomarker-driven studies.

Development and Application of Shiny Tools for Biomarker Analyses
Qinghua Song
Gilead Sciences, Inc.
qsong@gilead.com

Most biomarker analyses are exploratory in nature. Commonly produced biomarker exploratory analyses are either profiling the temporal change of biomarker after a drug intervention or studying the association of biomarker and other clinical or disease factors. Tables, figures, and listings (TFL) are costly to produce and time consuming to review with multiple iterations. With conventional programming practices, a large volume of TFL can be produced due to the iterative nature of data exploration and multiple dimensional data in biomarker research. For analysts, it is time consuming with redundant coding. For scientists, it is not convenient to review these outputs one by one. We built an interactive web-based tool for efficient data visualization and analysis reducing unnecessary iterations of explorations. We developed a system called ESTAR (Efficient Statistical Tool for Analysis and Reporting) to efficiently generate structured statistical reports with interactive components. Furthermore, we built advanced analytical tools for predictive biomarker evaluation and cut-point selection.

Session 18: Challenges in the Analysis of Large Spatial Data.

Construction of Space-Time Matern Correlation Functions
Tonglin Zhang and Hao Zhang
Purdue University
elzhang@purdue.edu

We provide a way to construct nonseparable space-time Matern correlation function based on the method of characteristic functions. Spatial time Matern correlation function is the characteristic function a multivariate t-distribution. Using a multivariate t-distribution for the space and a univariate t-distribution for time, we construct the family of space-time Matern correlation functions. Nonseparable Matern correlation function is derived if the dependence between the two t-distribution is considered.

Bootstrap variance estimation for one-per-stratum spatial sampling design
Zhonglei Wang and Zhengyuan Zhu
Iowa State University
wangzl@iastate.edu

In areal sampling, one-per-stratum design is a common approach which can achieve spatial balance and improve the precision of the resulting estimators. The downside of such design is that it is more challenging to have good design based variance estimation. In this paper, we propose a general class of stratified sampling design which produces spatially balanced sample. The generated sample is used to get the M-estimator of the coefficients in a spatial linear regression model, and a resampling approach is used to obtain the corresponding variance estimate. Asymptotic properties of the M-estimator and resampling based variance estimator under the proposed design are studied. Simulation studies are conducted to show the spatial balance of the samples from the proposed design and validate the asymptotic results for the resampling method for the variance estimation.

The Big Data Issues in Spatial Statistics
Hao Zhang
Purdue University
zhanghao@purdue.edu

One of areas where big data are collected is in environmental and climate studies. The Global Circulation Models or Regional Circulation Models can generate huge amount of data in space and time. Data collected through remote sensing or sensor networks are also huge. All these data are correlated spatially and temporally. One therefore has to deal with the huge covariance matrix in the traditional likelihood-based inferences or Bayesian inferences. When the dimension is extremely large, inversion of the matrix becomes infeasible numerically and also unstable due to the ill condition of the matrix. In this talk, I will discuss some recent developments in the theory and methods for dealing with the big spatial data, and in particular highlight some numerical algorithms that can find the low-rank structure in the covariance matrix.

A Multivariate Spatial Modelling Approach with Non-parametric Cross-covariogram
Yong Wang1 and Hao Zhang2
1Ohio Northern University
2Purdue University
yiwang.4@onu.edu

Analyzing multivariate spatial data is a delicate issue which brings challenges to model building, parameter estimation and computation. Most existing multivariate spatial models place restrictions on the choice of univariate model for each of the marginal processes and the parameter space of the cross-covariance function, which limits their potential applicability, especially when the model is fitted to data of higher dimension. We propose a new multivariate spatial modeling approach which uses parametric marginal models and a non-parametric cross-covariance function. This approach allows the users the flexibility to choose any marginal model and can be easily extended to handle multivariate spatial data of higher dimension. We will also discuss some efforts to deal with potential computational issues with parameter estimation.

Session 19: New applications in survival and longitudinal data

Statistical inference on quantile residual life with clustered survival data
Yun Chan Chi, I-Wen Lan and Tsung-Hsien Tsai
National Cheng Kung University
ycchi@mail.ncku.edu.tw

Clustered survival data arise naturally from biomedicine, econometrics and sociology studies. It is useful to estimate quantile residual life for medical devices, such as ventilating tubes in ears, artificial joints, and dental implant. For example, the patients with joint replacement may want to know how long the artificial joints will still be functioning given the fact that the artificial joints are used for at least 7 years. For clustered right-censored data, to the best of our knowledge, there are no available confidence intervals for quantile residual lifetime of a single group. Moreover, failure to account for
intra-cluster correlation can lead to bias inference. Therefore, this paper will extend the score tests considered by Jeong et al. (2008) to clustered right-censored data. Moreover, a confidence interval for quantile residual lifetime is constructed based on a Wald type statistic where the variance of sample quantile residual life function is estimated by a length-based variance estimator (Tsai et al. 2016). Simulation studies are conducted to evaluate the coverage rates of the proposed confidence intervals.

**Multi-state Event Analysis with Dynamic Covariates, Time-varying Coefficients and Measurement Errors**

Chuoxin Ma and Jianxin Pan  
School of Mathematics, University of Manchester  
[chuoxin.ma@manchester.ac.uk](mailto:chuoxin.ma@manchester.ac.uk)

A class of multi-state models with dynamic covariates are proposed to investigate the mechanism that generates the time-to-event data. The event flow consists of a series of recurrent events and each is followed possibly by a terminal event. Stratified Andersen-Gill models are assumed for the transition intensities. Among the covariates, some are dynamic in the sense that they contain past information, particularly expressed as functions of previous recurrent event occurrences. This may shed light on how effects of other covariates are mediated through event history. Great flexibility is allowed in the proposed modelling framework as time-fixed and time-dependent coefficients can be addressed in the same intensity function. Covariates that are exactly observed and those with measurement errors can also be handled in the same model. To estimate the parameters and unknown functions, local corrected score estimating equations together with one-step backfitting algorithm is employed. We evaluate the performance of our proposed models and estimation procedure through simulation studies and demonstrate its feasibility by applying it to a primary biliary cirrhosis (PBC) dataset.

**Comparison of different approaches for dynamic prediction of survival using longitudinal data**

Xuelin Huang  
University of Texas MD Anderson Cancer Center  
xihuang@mdanderson.org

Dynamic prediction is an important statistical tool for aiding medical decision-making, such as early detection of disease onset, post-treatment monitoring of disease prognosis. For this purpose, subjects’ biomarker values are repeatedly measured over time during follow-up visits. Predictions are conducted on a real-time basis so that at any time during follow-up, as soon as a new biomarker value is obtained, the prediction can be updated immediately to reflect the latest prognosis. Longitudinal biomarker trajectories are usually not linear, or even not monotone, and vary greatly across individuals. Therefore, it is difficult to fit them by parametric models. In this talk, I will first review the commonly used approaches for dynamic prediction, such as landmark analysis and joint modeling of longitudinal and survival data. Then I will introduce some of my recent work with my colleagues to enrich the current methods. These include quantile regression on residual survival time, functional principal component analysis for summarizing the changing patterns of patients’ longitudinal biomarker trajectories, and a supermodel to smoothly extend landmark analyses on discrete time points to the whole follow-up time interval. Simulation studies show that the proposed approaches achieve stable estimation of biomarker effects over time, and are robust against model misspecification. Moreover, they have better predictive performance than current methods, as evaluated by the root mean square error and area under the curve of receiver’s operating characteristics. The proposed methods are applied to a data set of patients with chronic myeloid leukemia. At any time following their treatment with tyrosine kinase inhibitors, longitudinally measured transcript levels of the oncogene BCR-ABL are used to predict patients’ risks of disease progression.

**Statistical Monitoring of Clinical Trials with Semi Competing Risks Outcomes**

Toshiimitsu Hamasaki1, Tomoyuki Sugimoto2, Scott R. Evans3 and Koko Asakura1  
1National Cerebral and Cardiovascular Center  
2Kagoshima University  
3Harvard T.H. Chan School of Public Health  
toshi.hamasaki@ncvc.go.jp

In this presentation, we discuss logrank test-based group-sequential methods for monitoring two event-time outcomes in the two typical situations as seen in cardiovascular disease or oncology trials: (a) both events are non-composite and one event is fatal, and (b) one event is composite and other event is fatal and a component of the composite event. In order to construct an efficient group-sequential design in such situations, we investigate operating characteristics of several monitoring strategies for early efficacy stopping in the two scenarios, where a trial is designed to: (1) evaluate effects for all of the outcomes (i.e., multiple co-primary endpoints: MCPE) or (2) evaluated effects for at least one outcome (i.e., multiple primary endpoints: MPE). Our main questions in this paper are whether both or either of event-time outcomes should be monitored during a trial and what is a better strategy for early efficacy stopping in terms of efficiency in terms of power, sample sizes and event numbers.

**Session 20: Recent advances in statistical analysis of genetic/genomic data**

**Testing variance components in the presence of nuisance boundary parameters**

Fei Zou  
University of Florida  
fayzou@ufl.edu

Mixed effects models are of great importance in genetic data analysis, and parameters of interest are often tested with likelihood ratio tests (LRT). When these parameters are variance components, the problem is considered non-standard as the parameters lie on the boundary of the parameter space. In certain settings, the asymptotic null distribution of the LRT is a mixture of chi-squared random variables, but is more complicated when there are nuisance boundary parameters. Using the incorrect reference distribution (as is often done in practice) can result in erroneous inference. In addition, existing asymptotic results are sensitive to the subtle assumption that the response vector must be able to be divided into independent units under both the null and alternative hypotheses. In an interesting dataset generated from recombinant inbred crosses of mice derived in the Collaborative Cross, when this independence assumption is violated, we show that commonly used reference distributions can perform quite badly. Alternative solutions will be discussed in this talk.

**A group LASSO based approach for GWAS meta-analysis across platforms**

Chi Song1, Yuan Jiang2, Kai Li2 and Shuyuan Lou3  
1Ohio State University  
2Oregon State University  
songj.1188@osu.edu
Since the wide application of SNP array and next generation sequencing technology, numerous genome wide association studies (GWAS) have been accumulated in the public domain such as dbGaP. Therefore, meta-analysis becomes an appealing approach to combine multiple GWASs that target on the same trait. However, GWAS meta-analysis often involves studies using different genotyping platforms, which can be solved by imputing the genotype data in each platform to the 1000 Genome Project or HapMap reference. This imputation step is time-consuming and calls for extra caution in the next analysis. In addition, because the studies combined may come from different populations with different linkage disequilibrium structures, we cannot assume the effect sizes of the SNPs stay the same across studies. To avoid these issues, we will propose a group LASSO based meta-analysis approach which requires no imputation. Specifically, we group the SNPs into their nearby genes and apply the group LASSO method to select SNPs belongs to the same gene in the combined studies. We evaluate the performance of this method using both simulation and real example in schizophrenia.

A semi-parametric test using multiple candidate kernels for gene set association analysis

* Tao He1, Shaoyu Li2, Ping-Shou Zhong3 and Yuehua Cai4
1 San Francisco State University
2 University of North Carolina at Charlotte
3 Michigan State University
hetao@sfsu.edu

Single-variant based genome-wide association studies have successfully detected many genetic variants that are associated with many complex traits. However, their power is limited due to weak marginal signals and ignoring potential complex interactions among genetic variants. Set-based strategy was proposed to provide a remedy, where multiple genetic variants in a given set (e.g., gene or pathway) are tested jointly so that systematic effect of the set is considered. In this work, we propose an efficient testing procedure that can not only control type I error rate but also generate power close to the one obtained using best kernel among all candidates. Our method is built upon the kernel-testing framework and is based on asymptotic results under a high-dimensional setting. Hence it can efficiently deal with the case where the number of variants in a set is larger than the sample size. Both simulation and real data analysis demonstrate the advantages of the method.

Integrating Transcriptional Time Lag Information into Gene Regulatory Network Construction through O

* Yaqun Wang1, Runze Li2 and Rongling Wu2
1 Rutgers, The State University of New Jersey
2 The Pennsylvania State University
yw305@sph.rutgers.edu

Using gene expression data to reconstruct Gene regulatory networks (GRN) can give new insights into the regulatory mechanisms underlying biological functions and phenotypic characteristics of an organism. Ordinary Differential Equation (ODE) modelling has been successfully used for inference of GRN, but this approach usually captures only instant regulation effects between genes. However, we know from biological knowledge that it takes time for regulator genes to have regulation impact on the targets. It may lead to misspecified regulation effects if only instant impacts are considered when reconstructing GRN. We address this issue by integrating the information of transcriptional time lags into ODE models to improve the accuracy of the GRN reconstruction. A procedure for GRN inference is developed with three steps including clustering genes, deciding potential regulators with transcriptional time lags and detecting regulation effects. A functional clustering approach based on Legendre Orthogonal Polynomial is applied in the first step for grouping genes according to the similarity of expression profiles over time. The model provides an unprecedented tool to help ones to understand a comprehensive picture of GRN. It has been well demonstrated by analysing real data sets and through extensive simulation studies.

Session 21: Title of session: Bayesian methods in Biostatistics

Bayesian predictive modeling for genomic-based personalized treatment selection

Jiansheng Ma1, Francesco Stingo2 and Brian Hobbs3
1 University of Texas, MD Anderson Cancer Center
2 University of Forence
bphobbs@mdanderson.org

Cancer is a complex dynamic microevolutionary process. Treatment requires understanding of the alterations within cell signaling pathways that enable cancer cells to evade cell death, proliferate, and migrate. Yet, the extent of variation among cancer cells within the same tumor make the disease inherently heterogeneous. The study individual candidate genes, signaling pathways, behaviors, or environmental exposures has limited our understanding of many areas of oncology. Future breakthroughs in personalized medicine will rely on molecular signatures that derive from synthesis of multifarious interdependent molecular quantities requiring more advanced quantitative methods. Moreover, because each tumor is unique, patients should not be considered exchangeable statistically. Rather, the extent to which results from previously treated patients inform our expectation of treatment success for a future patient should depend upon our current understanding of the extent to which each new patient’s tumor exhibits similarity with those previously treated. In this article, we introduce a Bayesian predictive framework that enables personalized treatment selection for new patients based on the treatment histories and molecular measurements of previously treated patients. We evaluate the method’s capacity to learn from accumulating information in the presence of complex predictive relationships between molecular quantities and treatments by implementing sequential treatment assignments using a well-known data set of leukemia (Golub et al., 1999; Brunet et al., 2004).

Bayesian Methods for Evaluating Air Quality Policies

Corwin Zigler
Harvard TH Chan School of Public Health
czigler@hsph.harvard.edu

This talk outlines methods to evaluate the extent to which the effect of a power plant regulatory intervention on air pollution is mediated through effects on power plant emissions. Power plants emit various compounds that contribute to ambient pollution, necessitating new methods to accommodate multiple interacting factors that are measured contemporaneously. We leverage two related frameworks for causal inference in the presence of mediating variables: principal stratification and causal mediation analysis. Both approaches are anchored to the exact same model for the observed data, which we specify with flexible Bayesian nonparametric techniques. The principal stratification and causal mediation analyses are interpreted in tandem to provide the first empirical investigation of the presumed causal pathways that motivate a variety of air quality regulatory policies.
A Bayesian Non-Parametric Causal Inference Model for Comparative Effectiveness Research
Chenguang Wang and Gary Rosner
Johns Hopkins University
cwang68@jhmi.edu

Comparative effectiveness research (CER) is designed to synthesize evidence of the benefits and harms of a treatment option from disparate sources including randomized clinical trials, observational studies and registry studies. The task of addressing study-specific heterogeneities is one of the most difficult challenges in CER. Bayesian hierarchical model with non-parametric extension provide a powerful and convenient platform that formalizes the information borrowing strength across the studies. In this paper, we propose a propensity score-based Bayesian non-parametric Dirichlet process mixture model that summarizes information from multiple observational and randomized studies to draw inference on the causal treatment effect. Simulation studies are conducted to evaluate the model performance under different scenarios.

Bayesian Nonparametric Approaches to Causal Inference on Quantiles
Dandan Xu1, Micheal Daniels3 and Almut Winterstein2
1The University of Texas at Austin
2University of Florida
Xudandanpuck@gmail.com

We propose methods for causal inference on quantiles using a Bayesian nonparametric (BNP) approach in the presence of many confounders. In particular, we define relevant causal quantities and specify BNP models to avoid bias from restrictive parametric assumptions. We first use Bayesian additive regression trees (BART) to model the propensity score and then construct the distribution of potential outcomes given the propensity score using a Dirichlet process mixture (DPM) of normals model. We thoroughly evaluate the operating characteristics of our approach and compare it to Bayesian and frequentist competitors. We use our approach to answer an important clinical question involving acute kidney injury using electronic health records.

Session 22: Causal inference with experimental and observational data

Instrumental variables as bias amplifiers with general outcome and confounding
Peng Ding
University of California, Berkeley
pengdingpku@berkeley.edu

Drawing causal inference with observational studies is the central pillar of many disciplines. One sufficient condition for identifying the causal effect is that the treatment-outcome relationship is unconfounded conditional on the observed covariates. It is often believed that the more covariates we condition on, the more plausible this unconfoundedness assumption is. This belief has had a huge impact on practical causal inference, suggesting that we should adjust for all pretreatment covariates. However, when there is unmeasured confounding between the treatment and outcome, estimators adjusting for some pretreatment covariate might have greater bias than estimators without adjusting for this covariate. This kind of covariate is called a bias amplifier, and includes instrumental variables that are independent of the confounder, and affect the outcome only through the treatment. Previously, theoretical results for this phenomenon have been established only for linear models. We fill in this gap in the literature by providing a general theory, showing that this phenomenon happens under a wide class of models satisfying certain monotonicity assumptions. We further show that when the treatment follows an additive or multiplicative model conditional on the instrumental variable and the confounder, these monotonicity assumptions can be interpreted as the signs of the arrows of the causal diagrams.

Bayesian regression tree models for causal inference
Richard Hahn1, Jared Murray2 and Carlos Carvalho3
1University of Chicago
2Carnegie Mellon University
3University of Texas
richard.hahn@chicagobooth.edu

This paper develops a semi-parametric Bayesian regression model for estimating heterogeneous treatment effects from observational data. Standard nonlinear regression models, which may work quite well for prediction, can yield badly biased estimates of treatment effects when fit to data with strong confounding. Our Bayesian causal forests model avoids this problem by directly incorporating an estimate of the propensity score function in the specification of the response model, implicitly inducing a covariate-dependent prior on the regression function. This new parametrization also allows treatment heterogeneity to be regularized separately from the prognostic effect of control variables, making it possible to informatively “shrink to homogeneity”, in contrast to existing Bayesian non- and semi-parametric approaches.

Trustworthy Analysis of Online A/B Tests: Pitfalls, challenges and solutions
Alex Deng, Jiannan Lu and Jonathan Litz
Microsoft Corporation
jiannli@microsoft.com

A/B tests (or randomized controlled experiments) play an integral role in the research and development cycles of technology companies. As in classic randomized experiments (e.g., clinical trials), the underlying statistical analysis of A/B tests is based on assuming the randomization unit is independent and identically distributed (i.i.d.). However, the randomization mechanisms utilized in online A/B tests can be quite complex and may render this assumption invalid. Analysis that unjustifiably relies on this assumption can yield untrustworthy results and lead to incorrect conclusions. Motivated by challenging problems arising from actual online experiments, we propose a new method of variance estimation that relies only on practically plausible assumptions, is directly applicable to a wide range of randomization mechanisms, and can be implemented easily. We examine its performance and illustrate its advantages over two commonly used methods of variance estimation on both simulated and empirical datasets. Our results lead to a deeper understanding of the conditions under which the randomization unit can be treated as i.i.d. In particular, we show that for purposes of variance estimation, the randomization unit can be approximated as i.i.d. when the individual treatment effect variation is small; however, this approximation can lead to variance underestimation when the individual treatment effect variation is large.

Using Survival Information in Truncation by Death Problems Without the Monotonicity Assumption
Fan Yang1 and Peng Ding2
1University of Chicago
2University of California, Berkeley
fyang@health.bsd.uchicago.edu

A/B tests can be quite complex and may render this assumption invalid. Analysis that unjustifiably relies on this assumption can yield untrustworthy results and lead to incorrect conclusions. Motivated by challenging problems arising from actual online experiments, we propose a new method of variance estimation that relies only on practically plausible assumptions, is directly applicable to a wide range of randomization mechanisms, and can be implemented easily. We examine its performance and illustrate its advantages over two commonly used methods of variance estimation on both simulated and empirical datasets. Our results lead to a deeper understanding of the conditions under which the randomization unit can be treated as i.i.d. In particular, we show that for purposes of variance estimation, the randomization unit can be approximated as i.i.d. when the individual treatment effect variation is small; however, this approximation can lead to variance underestimation when the individual treatment effect variation is large.

Using Survival Information in Truncation by Death Problems Without the Monotonicity Assumption
Fan Yang1 and Peng Ding2
1University of Chicago
2University of California, Berkeley
fyang@health.bsd.uchicago.edu

In some randomized clinical trials, patients may die before the measurements of their outcomes. Even though randomization generates comparable treatment and control groups, the remaining survivors often differ significantly in background variables that are prognostic to the outcomes. This is called the truncation by death problem. Under the potential outcomes framework, the only well defined causal effect on the outcome is within the subgroup of patients who would always survive under both treatment and control. Because the definition of the subgroup depends on the potential values of the survival status that could not be observed jointly, without making strong parametric assumptions, we cannot identify the causal effect of interest and consequently can only obtain bounds of it. Unfortunately, however, many bounds are too wide to be useful. We propose to use detailed survival information before and after the measurements of the outcomes to sharpen the bounds of the subgroup causal effect. Because survival times contain useful information about the final outcome, carefully utilizing them could improve statistical inference without imposing strong parametric assumptions. Moreover, we propose to use a copula model to relax the commonly invoked but often doubtful monotonicity assumption that the treatment extends the survival time for all patients.

Session 23: Recent Developments in Early Phase Dose-Finding

Sequential Quantile Estimation Using Continuous Outcomes:
With Applications in Dose Finding
♦ Chih-Chi Hu and Ying Kuen Cheung
1 Sanofi
2 Columbia University
chih-chi.hu@sanofi.com

 Typically, phase I trials are designed to determine the maximum tolerated dose, defined as the maximum test dose that causes a toxicity with a target probability. In this talk, we formulate dose finding as a quantile estimation problem and focus on situations where toxicity is defined by dichotomizing a continuous outcome. While the majority of existing dose-finding designs work with dichotomized data, substantial information may be lost due to dichotomization. In this light, there is a strong reason for using the original continuous outcome. We introduce least square recursion, a versatile method for sequential quantile estimation using continuous data. Simulations in the context of a real phase I trial show that, compared to a method that utilizes dichotomized data, least square recursion recommends the optimal dose more often without exposing more patients to unsafe doses.

Sample size determination for the two-stage continual reassessment method (CRM)
♦ Cody Chiu, Zilan Chai and Ken Cheung
Columbia University
cc3780@columbia.edu

In the last few years, various model-based designs have been developed to identify the maximum tolerated dose (MTD) in phase I trials. Encouragingly, an increasing number of trials have started to implement some of these novel methods, amongst which the Bayesian Continual Reassessment Method (CRM) is the most widely used. The CRM proposes two main dose-finding strategies: one-stage design and two-stage design. The latter requires the specification of an initial design (e.g., “3+3”) that stays in effect until the first observed toxicity; at that point the trial turns to the model-based CRM. The two-stage CRM represents a more conservative escalation alternative by starting the trial at the lowest dose and potentially reducing the risk of overdosing. In order to facilitate the planning stage of a trial and provide a comparison of the two designs, we extended the sample size formulae for the one-stage CRM (Cheung, 2013) to the two-stage design. Simulation results are presented for several scenarios by varying parameters such as: targeted toxicity and accuracy probabilities, number of pre-specified dose levels and fixed cohort sizes used in the first (algorithmic) part of the trial.

Parametric Dose Standardization for Two-Agent Phase I-II Trials with Ordinal Efficacy and Toxicity

Peter Thall
M.D. Anderson Cancer Center
rex@mdanderson.org

A Bayesian model and design are described for a phase I-II trial to jointly optimize the doses of two agents based on 4-level ordinal efficacy and toxicity. To reflect possibly complex joint effects of the two doses on each outcome, in each marginal distribution a generalized continuation ratio model is assumed, with each agent’s dose parametrically standardized in the linear term. Elicited numerical utilities of the 16 elementary outcomes are used to compute posterior mean utilities as criteria for selecting dose pairs. Adaptive randomization is used to reduce the risk of getting stuck at a sub-optimal dose pair. A simulation study shows that parametric dose standardization with additive dose effects provides a robust model for choosing dose pairs in this setting, and it compares favorably with designs based on alternative models having conventional dose-dose interactions.

Session 24: Recent Advances in Missing Data Methods

Survival Analysis with Presence of Informative Censoring via Nonparametric Multiple Imputation
♦ Chengcheng Hu, Jeremy Taylor and Chiu-Hsieh Hsu
1 University of Arizona
2 University of Michigan
hucc@email.arizona.edu

We propose a nonparametric multiple imputation approach to recover information for censored observations while analyzing survival data with presence of informative censoring. A working shared frailty model is proposed to estimate the magnitude of informative censoring through measuring Kendall’s tau, which is only used to determine the size of imputing risk set for each censored subject. Specifically, a larger tau indicates a smaller size of the imputing risk set. We have shown that the distance between the posterior means of frailty is equivalent to the distance between the observed times. We, therefore, propose to use the observed times for subjects at risk to calculate the distance from each censored subject and combine it with tau to select an imputing risk set for each censored subject. In simulation, we have shown that the nonparametric multiple imputation approach produces survival estimates comparable to the targeted values and coverage rates of the confidence intervals comparable to the nominal level even in a situation with a high degree of informative censoring.

Variable selection in the presence of missing data: resampling and imputation
Qi Long, Changgee Chang and Brent Johnson
1 University of Pennsylvania
2 University of Rochester
changgee@mail.med.upenn.edu
In the presence of missing data, variable selection methods need to be tailored to missing data mechanisms and statistical approaches used for handling missing data. We focus on the mechanism of missing at random and variable selection methods that can be combined with imputation. We investigate a general resampling approach (BI-SS) that combines bootstrap imputation and stability selection, the latter of which was developed for fully observed data. The proposed approach is general and can be applied to a wide range of settings. Our extensive simulation studies demonstrate that the performance of BI-SS is the best or close to the best and is relatively insensitive to tuning parameter values in terms of variable selection, compared with several existing methods for both low-dimensional and high-dimensional problems. The proposed approach is further illustrated using two applications, one for a low-dimensional problem and the other for a high-dimensional problem.

Semiparametric pseudoscore for regression with multidimensional but incompletely observed regressor

Zonghui Hu, Jing Qin and Dean Follmann
National Institutes of Health
huzo@niaid.nih.gov

We study the regression with incomplete observation in the regressors. Let Y stand for the response, Z the vector of fully observed regressors, and X the regressor with incomplete observation. To handle missing data, maximum likelihood estimation via expectation-maximisation (EM) is the most efficient but is sensitive to the specification of the distribution of the incompletely observed regressor. Under a missing at random assumption, we propose an EM type estimation via a semiparametric pseudo-score. Like in EM, we derive the conditional expectation of the score function given Y and Z, or the mean score, over the incompletely observed units under a postulated distribution of X. Instead of directly using the “mean score” in estimating equation, we use it as a working index to construct the semiparametric pseudo-score via nonparametric regression. Introduction of semiparametric pseudo-score into the EM framework reduces sensitivity to the specified distribution of X. It also avoids the curse of dimensionality when Z is multidimensional. The resulting regression estimator is more than doubly robust: it is consistent if either the pattern of missingness is correctly specified or the working regression estimator is more than doubly robust; it is consistent to leaving an O(n−1/2) subset of the sample that ought to have been removed. Accordingly, our solution uses bounded-influence regression methods to adjust for the running variable, these methods having been identified by statistical theory as being uniquely insensitive to sample contamination (He, 1991; Yohai and Zamar, 1997).

Session 25: Exposure Response Modeling

Statistical Considerations in Exposure Response Modeling

Guoxing Song
FDA
guoxing.song@fda.hhs.gov

An understanding of exposure-response relationships linking dose, concentration and response is an important aspect in regulatory decision-making. Examples of the use of exposure-response relationships range from providing supportive or confirmatory evidence for approval to guiding dose adjustments for a subpopulation. Ideally, the concentration-response relationship in the same individual over time is most desired as this relationship is not confounded by dose selection/titration phenomena and inter-individual PK variability. However the concentration-response relationship may be established via modeling across different concentrations and different subjects potentially resulting in confounding. In this talk, we will discuss statistical validity and issues related to the design and analysis of exposure-response relationships. We will also highlight applications through examples.

Assay Sensitivity Analysis using Exposure Response Modeling in “Hybrid TQT” Study

Dalong Huang, Janell Chen, Qianyu Dang and Yi Tsong
FDA/CDER
dalong.huang@fda.hhs.gov

The ICH E14 guidance requires drug sponsors to complete a thorough QT/QTc (TQT) study? to evaluate the effect of a drug on cardiac repolarization. This guidance was revised in December 2015. It now enables the use of exposure response (ER) modeling as primary analysis to provide definitive evidence of the lack of a QT effect of a drug in development. Since the revision of ICH E14 guidance, some pharmaceutical companies have started to use the hybrid TQT study to meet ICH E14 regulatory requirement. The hybrid TQT study includes same treatment arms (therapeutic and/or supratherapeutic dose of investigational drug, placebo, and positive control) with approximately half the sample size as traditional TQT study, but use ER as primary analysis and assay sensitivity analysis.
Recently, much attention has been received on the analysis of functional imaging data to delineate the intrinsic functional connectivity pattern among different brain regions within each subject. However, only few approaches for integrating functional connectivity pattern from multiple subjects have been proposed. The goal of this study is to develop a reduced-rank model framework for analyzing the whole-brain voxel-wise functional images across multiple subjects in the frequency domain. Considering the neighboring voxels with different weights, the frequency and spatial factors can be extracted. Imposing sparsity on the frequency factors enables us to identify the dominant frequencies. In addition, the spatial maps can be used for detecting group difference, when the comparison between different groups is of specific interest. Simulation study shows that the proposed method achieves less spatial variability and better estimates of frequency and spatial factors, compared to some existing methods. Finally, we apply the proposed method to ADNI data.

Estimation of Anatomically Informed Functional Networks via Bayesian Gaussian Graphical Model

Lxavier A. Higgins1, Suprateek Kundu2, *Subhadip Pal3 and Ying Guo2

1University of Louisville
2Emory University
3University of Texas MD Anderson Cancer

mahn@unr.edu

Recently, much attention has been received on the analysis of functional imaging data to delineate the intrinsic functional connectivity pattern among different brain regions within each subject. However, only few approaches for integrating functional connectivity pattern from multiple subjects have been proposed. The goal of this study is to develop a reduced-rank model framework for analyzing the whole-brain voxel-wise functional images across multiple subjects in the frequency domain. Considering the neighboring voxels with different weights, the frequency and spatial factors can be extracted. Imposing sparsity on the frequency factors enables us to identify the dominant frequencies. In addition, the spatial maps can be used for detecting group difference, when the comparison between different groups is of specific interest. Simulation study shows that the proposed method achieves less spatial variability and better estimates of frequency and spatial factors, compared to some existing methods. Finally, we apply the proposed method to ADNI data.

Estimation of Anatomically Informed Functional Networks via Bayesian Gaussian Graphical Model

Lxavier A. Higgins1, Suprateek Kundu2, *Subhadip Pal3 and Ying Guo2

1University of Louisville
2Emory University
3University of Texas MD Anderson Cancer

mahn@unr.edu

Recently, much attention has been received on the analysis of functional imaging data to delineate the intrinsic functional connectivity pattern among different brain regions within each subject. However, only few approaches for integrating functional connectivity pattern from multiple subjects have been proposed. The goal of this study is to develop a reduced-rank model framework for analyzing the whole-brain voxel-wise functional images across multiple subjects in the frequency domain. Considering the neighboring voxels with different weights, the frequency and spatial factors can be extracted. Imposing sparsity on the frequency factors enables us to identify the dominant frequencies. In addition, the spatial maps can be used for detecting group difference, when the comparison between different groups is of specific interest. Simulation study shows that the proposed method achieves less spatial variability and better estimates of frequency and spatial factors, compared to some existing methods. Finally, we apply the proposed method to ADNI data.

Estimation of Anatomically Informed Functional Networks via Bayesian Gaussian Graphical Model

Lxavier A. Higgins1, Suprateek Kundu2, *Subhadip Pal3 and Ying Guo2

1University of Louisville
2Emory University
3University of Texas MD Anderson Cancer

mahn@unr.edu

Recently, much attention has been received on the analysis of functional imaging data to delineate the intrinsic functional connectivity pattern among different brain regions within each subject. However, only few approaches for integrating functional connectivity pattern from multiple subjects have been proposed. The goal of this study is to develop a reduced-rank model framework for analyzing the whole-brain voxel-wise functional images across multiple subjects in the frequency domain. Considering the neighboring voxels with different weights, the frequency and spatial factors can be extracted. Imposing sparsity on the frequency factors enables us to identify the dominant frequencies. In addition, the spatial maps can be used for detecting group difference, when the comparison between different groups is of specific interest. Simulation study shows that the proposed method achieves less spatial variability and better estimates of frequency and spatial factors, compared to some existing methods. Finally, we apply the proposed method to ADNI data.
In cognitive aging, cognition is optimally quantified over multiple subdomains called reference abilities: reasoning, speed processing, memory and vocabulary. In this study, we aim to identify reference ability-specific resting functional networks. We first identified resting state networks that are also activated or deactivated during cognitive tasks (Smith et al., 2009). Sparse functional multivariate multiple regression analysis was applied to derive parsimonious representation of reference ability specific resting state network patterns. We further tested age-moderation on the RA-specific resting functional networks representation.

Session 27: Advancing Cancer Clinical Trials

Suspension of accrual in phase II cancer clinical trials

Yimei Li1, Rosemarie Mick1 and Daniel Heitjan2
1University of Pennsylvania
2Southern Methodist University

Phase II cancer clinical trial designs commonly incorporate an interim analysis for lack of efficacy. To strictly and ethically implement such designs, one should suspend accrual in cases where pending patient outcomes can affect early termination decisions. This project aims to evaluate various options for accrual suspension and illustrate how the suspension strategy affects operating characteristics of the trial. We define a strict suspension strategy for determining whether one should continue, suspend, or restart accrual at any point within the trial. The strategy is compared to a naive implementation of suspension and a strategy of no suspension. We evaluate the methods’ operating characteristics by simulation. Results suggest that the suspension strategy has little effect on type I error, power, and early termination probability. Methods that involve stricter suspension policies generally lead to smaller but longer trials. Differences across strategies are substantial when the ratio of enrollment rate to outcome availability rate is high. We conclude that he suspension strategy is most relevant in trials that accrue rapidly and require lengthy observation of each subject. The choice of suspension strategy involves a tradeoff between the cost of implementing a potentially complex suspension algorithm in real time.

Dose finding designs for late-onset and cumulative toxicities

Shing Lee1, Moreno Ursino2, Ying Kuen Cheung1 and Sarah Zohar2
1Columbia University
2INSERM

The toxicity profile of newer anticancer treatments such as targeted and immunotherapeutic agents differs from that of chemotherapy. While some of these newer agents cause dose limiting toxicities, others cause lower grade toxicities which may not occur within the first cycle of treatment. Thus, in the early development of these agents it is necessary to account for both lower grade toxicity, as well as, late onset and cumulative toxicity. However, methods for ordinal toxicity that can account for lower grade toxicities require complete follow-up data from patients before the next dose can be assigned. If the follow-up time to include late onset and cumulative toxicities is long, it will have an impact on time to completion of these studies and delay middle and late development. We propose an extension to the continual reassessment method with multiple constraints that can accommodate for incomplete follow-up data. This method allows for patients to be entered before complete follow-up is observed and can impose constraints on milder toxicities. We present the method in the context of a targeted therapy in patients with lymphoma.

A multi-stage adaptive enrichment design to improve efficiency of early phase clinical trials

Brandon Luber and Hao Wang

Clinical trials in the early stage of oncology drug development need to be more efficient in targeting those patients who will most likely benefit from treatment, especially as we move into costly late-stage development. It is also crucial for the design of early phase trials to take advantage of our understanding of molecular pathology of tumors and the mechanism of the drug. With this in mind, we present an adaptive design strategy that incorporates both molecular assessment and subpopulation enrichment. We have implemented the strategy for a phase II prostate cancer study. Olaparib, a poly-ADP ribose polymerase (PARP) inhibitor, has demonstrated preliminary efficacy in metastatic castration-resistant prostate cancer, most notably in patients with mutations in DNA repair genes. In a trial to test Olaparib in earlier disease states, we have developed a novel multi-stage adaptive enrichment design. This single arm trial will enroll up to 50 patients, with interim stopping rules to determine futility or need for enrichment of the study population. Tumors are sequenced for mutations in DNA repair genes at interim analysis, and the enriched population will be determined based on their deleterious mutations. Simulations demonstrate that the design achieves adequate power to detect an effect in the unsampled population and in the selected population (if there is no effect in the unselected) while maintaining type I error control.

Design considerations in immunotherapy trials with multiple targets and indications

Kay See Tan
Memorial Sloan Kettering Cancer Center

tank@mskcc.org

Major strides have been made in the development of targeted cancer therapy, but optimizing patient recruitment and selection for trials with multiple targets and indications remains an area of investigation. For example, targeted therapies are being developed for the most common mutation in lung adenocarcinoma. This mutation is frequently associated with overexpression of mesothelin (MSLN) and CA125 on tumor cell surface, offering ideal target antigens for T-cell therapy. These antigens are also overexpressed in other diseases such as breast and ovarian cancers, providing targets that extend beyond lung adenocarcinoma. Master/integrated protocols can streamline the development process of T-cell therapies in such a context to facilitate the transition from dose-finding to efficacy trials of combination therapies targeting both MSLN and CA125. However, most master protocols have been designed as parallel trials of subpopulations according to disease-target-drug combinations, and few take advantage of the correlated information that can be shared across the subpopulations. Using the motivating example, this talk will present potential master protocol designs and identify recently proposed statistical considerations (e.g., bridging dose-finding studies and basket aggregation) applicable at various points of the integrated trial structure to improve patient selection, minimize number of patients needed, and control family-wise error rate. Special considerations will also be given to issues pertinent to immunotherapy trials, such as on-target off-tumor toxicities.
Session 28: Recent Advance in Platform Clinical Trials in Era of Precision Medicine

A Subgroup Cluster Based Bayesian Adaptive Design for Precision Medicine

Wentian Guo¹, Yuan Ji¹ and Daniel Catenacci²
¹Laiya Consulting, Inc.
²University of Chicago
wentianguo@laiyaconsulting.com

In precision medicine, a patient is treated with targeted therapies that are predicted to be effective based on the patient’s baseline characteristics such as biomarker profiles. Oftentimes, patient subgroups are unknown and must be learned through inference using observed data. We present SCUBA, a Subgroup CIUster based Bayesian Adaptive design aiming to fulfill two simultaneous goals in a clinical trial, 1) to enrich the allocation of each subgroup of patients to their precision and desirable treatments and 2) to report multiple subgroup-treatment pairs (STPs). Using random partitions and semiparametric Bayesian models, SCUBA provides coherent and probabilistic assessment of potential patient subgroups and their associated targeted therapies. Each STP can then be used for future confirmatory studies for regulatory approval. Through extensive simulation studies, we present an application of SCUBA to an innovative clinical trial in gastroesophageal cancer.

Increasing Efficiency of Oncology Basket Trials Using a Bayesian Approach

Rong Liu
University of Toledo
rong.liu@utoledo.edu

With the rapid growth of targeted and immune-oncology therapies, novel statistical design approaches are needed to increase the flexibility and efficiency of early phase oncology trials. Basket trials enroll patients with defined biological deficiencies, but with multiple histologic tumor types (or indications), to discover in which indications the drug is active. In such designs different indications are typically analyzed independently. This, however, ignores potential biological similarities among the indications. Our research provides a statistical methodology to enhance such basket trials by assessing the homogeneity of the response rates among indications at an interim analysis, and applying a Bayesian hierarchical modeling approach in the second stage if the efficacy is deemed reasonably homogeneous across indications. This increases the power of the study by allowing indications with similar response rates to borrow information from each other. Via simulations, we quantify the efficiency gain of our proposed approach relative to the conventional parallel approach. The operating characteristics of our method depend on the similarity of the response rates between the different indications. If the response rates are comparable in most or all indications after treatment with the investigational drug, a substantial increase in efficiency as compared to the conventional approach can be obtained as fewer patients are required or a higher power is attained. A substantial variation in the response rates between the indications, however, is demonstrated to be associated with only a minimal loss of efficiency of our method.

Early Signal Detection in Basket Trial via Bayesian Approach

Charlie Cao and Wenwen Zhang
Takeda
charlie.cao@takeda.com

It is important to make early GO/NO GO decisions in clinical development to avoid waste of resources. The decision is not only to identify the promising drug candidates, but also to find the potential indications in order to speed up drug development process. Within Bayesian Framework, we can calculate the likelihood of a drug candidate for various endpoints and decide which indication should be pursued. Some motivating examples in CNS area and results based on simulation studies will be discussed.

Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials

Jianchang Lin¹ and Veronica Bunn²
¹Takeda Pharmaceuticals
²Florida State University
jianchang.lin@takeda.com

Platforms/Umbrella trials are clinical trials that allow for concurrent evaluations of multiple treatments in single master protocol, thus allowing for more efficient and ethical studies compared to traditional two-arm trials. There are few major types of design could be implemented in the platform trials: (1) Conventional group-sequential multi-arm multi-stage (MAMS) designs use pre-specified stopping boundaries and treatment selection rules to determine if experimental treatments should be dropped. Further more, flexible MAMS designs allow for interim modifications to the design plan without compromising error rates. (2) Bayesian response adaptive randomization (BRAR) designs increase patient allocation to treatment arms that are performing well during the course of the trial. In this paper, we compare these two major methods and their extensions under several scenarios in the platform trials setting. Results show that BRAR and flexible MAMS designs have comparable power and type 1 error rate under varying simulated scenarios, allowing for addition of flexible treatment selection. BRAR outperforms flexible MAMS when there is a single effective treatment. Flexible MAMS designs are more efficient compared to BRAR when there are no effective treatments. BRAR performance increases as the probability of a treatment arm being dropped increases. Examples of these platform trials application in oncology will also be demonstrated and discussed.

Session 29: Aspects of Statistical Inference

Tuning Parameter Selection in the LASSO with Unspecified Propensity

Jiwei Zhao
State University of New York at Buffalo
jiwei2012zhao@gmail.com

The least absolute shrinkage and selection operator (LASSO) is one of the most significant model selection methods introduced about two decades ago. Although it attracts numerous attentions in both theory and computation, we still encounter many difficulties in real applications. For instance, in a real data set, we may have various missing values. To correctly adopt the LASSO, we have to incorporate the missing data mechanism, or the propensity, in the penalized likelihood. Also, how to choose the tuning parameter is still an open problem. Especially with a “messy” data set. Two distinct contributions make our work different from all the existing literature. First, we allow the data set to have missing values by imposing a very general and flexible propensity. Compared to the missing data methods with a concrete propensity, this assumption is relatively easier to satisfy in reality and it makes our methodology more robust. Many existing ignorable or nonignorable missing data situations, as well as some biased sampling problems, belong to a special case under our consideration. Second, to determine the tuning parameter, we examine four different methods including cross validation (CV), Bayesian information criterion (BIC) and two others focusing on
estimation stability and variable selection stability. To illustrate our methods, we conduct comprehensive simulation studies and apply our methods to a real data for melanoma study.

**Multi-Panel Kendall Plot Applied to Measuring Dependence**  
Albert Vexler, •Georgios Afendras and Marianthi Markatou  
University at Buffalo  
gafendra@buffalo.edu

The Kendall plot (K-plot) is a plot measuring dependence between the components of a bivariate random variable. The K-plot graphs the Kendall distribution function against the distribution function of $VU$, where $V, U$ are independent $U[0,1]$ random variables. We associate K-plots with the receiver operating characteristic (ROC) curve, a well-accepted graphical tool in biostatistics for evaluating the ability of a biomarker to discriminate between two populations. In parallel with the area under the ROC curve, we propose a novel strategy to measure association between random variables from a continuous bivariate distribution. First, we discuss why the area under the Kendall curve (AUK) cannot be used as an index of dependence. We then suggest a simple and meaningful extension of the definition of the K-plots, and define an index of dependence that is based on AUK. This measure characterizes a wide range of two-variable relationships, thereby completely detecting the underlying dependence structure. Properties of the proposed index satisfy the mathematical definition of a measure. Finally, simulations and real data examples illustrate the applicability of the proposed method.

**Estimation of Two Ordered Normal Means when a Covariance Matrix is Known**  
•Watanabe(Chang) Genso(Yuan-Tsung)$^1$, Fukuda Kazufumi$^2$ and Shinozaki Nobuo$^3$  
$^1$Mejiro University  
$^2$Dept. of Pure and Applied Math. Waseda Univ.  
$^3$Keio University  
cho@mejiro.ac.jp

Estimation of two normal means with an order restriction is considered when a covariance matrix is known. It is shown that restricted maximum likelihood estimator (MLE) stochastically dominates both estimators proposed by Hwang and Peddada (1994) and Peddada et al. (2005). The estimators are also compared under the Pitman nearness criterion and it is show that the MLE is closer to ordered means than the other two estimators. Estimation of linear functions of ordered means is also considered and a necessary and sufficient condition on the coefficients is given for the MLE to dominate the other estimators in terms of mean squared error.

**Discrete Smoothing Kernels**  
Marianthi Markatou  
Dept. of Biostatistics, University at Buffalo  
markatou@buffalo.edu

Kernels are essential elements in the construction of learning systems and have received considerable attention in the machine learning literature. In statistics, kernels are used as tools for achieving specific data analytic goals such as density estimation or goodness of fit testing. We consider the problem of creating smoothing kernels for multinomial and product multinomial models. Our construction is based on the properties of continuous time Markov chains. We will discuss an algorithm for the construction of these kernels and exemplify its use in smoothing ordered categorical data. Furthermore, using these constructions, we will discuss independence tests that are analogues of the chi-squared test of independence.

**Session 30: Robustness Aspects of Optimal Experimental Design**

**Developments for Information-Based Subdata Selection**  
John Stufken  
Arizona State University  
jstufken@asu.edu

In order to do a statistical analysis of big data, due to size it may be necessary to perform the analysis on selected subdata. There are various methods for selecting subdata from big data, including sampling-based methods and methods that advocate the use of information-based criteria. The latter rely heavily on assumptions that, in many instances, can be questionable for big data. We will present the basic idea of information-based subdata analysis, discuss some of the recent extensions, and explore the sensitivity of the methods to assumptions.

**Robustness of Block Designs to Data Loss**  
JP Morgan  
Department of Statistics, Virginia Tech  
jpmorgan@vt.edu

Techniques and results for identifying block designs which are robust to loss of data are explored, with a focus on loss of entire blocks. Special attention is afforded the class of balanced incomplete block designs. The importance of combining theory and computation is emphasized.

**EW Optimality for Robust Experimental Design under Generalized Linear Models**  
Jie Yang  
University of Illinois at Chicago  
jieyangtong@gmail.com

Optimal designs under generalized linear models typically require assumed parameter values. For many applications, the experimenter may have little or limited information about the values of parameters. In that case, Bayes D-optimality which maximizes the expected log determinant of Fisher information matrix provides a reasonable solution for robust experimental designs. However, it suffers from heavy computational burden. In this talk, we introduce an alternative solution called EW D-optimality which maximizes the log determinant of expected Fisher information matrix instead. Thorough simulation studies across different models and choices of priors shows that EW D-optimal designs are much easier to calculate and still highly efficient compared with Bayes D-optimal designs. It provides a good surrogate for Bayes optimality towards robust experimental designs.

**Flexible Optimal Design Strategies**  
Timothy OBrien  
Loyola University Chicago  
tobriel@luc.edu

Researchers often find that nonlinear regression models are more applicable for modelling various biological, physical and chemical processes than are linear ones since they tend to fit the data well and since these models (and model parameters) are more scientifically meaningful. These researchers are thus often in a position of requiring optimal or near-optimal designs for a given nonlinear model. A common shortcoming of most optimal designs for nonlinear models used in practical settings, however, is that these designs typically focus only on (first-order) parameter variance or predicted variance, and thus ignore the inherent nonlinear of the assumed model function. Another shortcoming of optimal designs is that they often have only p support points, where p is the number of model parameters.
Abstracts

Measures of marginal curvature, first introduced in Clarke (1987) and further developed in Haines et al (2004), provide a useful means of assessing this nonlinearity. Other relevant developments are the second-order volume design criterion introduced in Hamilton and Watts (1985) and extended in O’Brien (2010), and the second-order MSE criterion developed and illustrated in Clarke and Haines (1995).

This talk examines various robust design criteria and those based on second-order (curvature) considerations. These techniques, coded in the GAUSS and SAS/IML software packages, are illustrated with several examples.

Session 31: Opportunities to innovatively leverage historical information to improve medical product development

Leveraging historical information in clinical programs
Freda Cooner
Sanofi
freda.cooner@sanofi.com

Historical information is often available for diseases and sometimes treatments of interest prior to the initiation of a clinical program. However, it is rarely used to support marketing applications, either for efficacy or safety of the investigational medical products. The obstacles include identifying accessible historical database, getting consensus from development team and the regulatory agencies, and developing a robust methodology to incorporate historical information into evidentiary support for the current clinical program. This presentation will illustrate the main sources of historical information, summarize the techniques using historical information at a trial design stage, and briefly discuss leveraging historical information in data analysis through Bayesian framework. Examples will be presented to display the difficulties in leveraging historical information and to elucidate the potential usage of Bayesian framework in clinical programs.

Bayesian methods toward a feasible and informative clinical trial in small populations
Margaret Gamalo
Eli Lilly & Co
gamalo.margaret@lilly.com

Developing drugs for a pediatric, orphan, or an unmet medical need indication is characterized by having a smaller population from which to recruit patients and presents a challenge toward designing a feasible, robust and informative trial. For this reason, most trials are designed with the preponderance that they lack sufficient ability to demonstrate efficacy in the traditional frequentist way. While this presents an ethical problem as it may violate the concept of scientific validity, this can possibly be mitigated by the application of innovative statistical methodologies, e.g., Bayesian approach. The approach permits a synthesis of information and provides an opportunity in reducing the sample size burden while still providing sufficient information about effects of treatments and their corresponding uncertainties. As an illustration, I will provide examples on the retrospective use of Bayesian methods in some clinical trials for rare indications in order to make a case for a broader use of the methodology in small population drug development.

Bayesian Historical Borrowing Method with Case Study in a Phase 3 HCV Trial
♦ Ran Liu1, Qi Tang2, Martin King1, Bo Fu1, Sandra Lovell1 and Alan Hartford2
1 AbbVie Inc.

A Bayesian framework to leverage historical control data in clinical trials is introduced. Under the proposed framework, data from historical trials of the control regimen are “borrowed” and combined with data from the control arm in current trial, which could lead to more accurate point estimates, greater power, and reduced type I error. The extent of borrowing is controlled by a Bayesian model which borrows more when the degree of similarity is larger between the control arms in the historical and new trials. A case study in a phase 3 trial in chronic hepatitis C virus (HCV) is described in detail to illustrate the motivation, procedure, and benefit of the method. Simulations show that borrowing historical control data meaningfully improves the power of the study while maintaining appropriate control of the type I error rate. The proposed method provides a means to incorporate data from historical trials and improves the efficiency of the current trial while maintaining appropriate control of the type I error rate.

Session 32: Student session (I)

Sequential Outcome-Weighted Multicategory Learning for Estimating Optimal Individualized Treatment
♦ Xuan Zhou1, Yuanjia Wang2 and Donglin Zeng1
1 University of North Carolina at Chapel Hill
2 Columbia University
xuanz@live.unc.edu

Personalized medicine has received increasing interest among clinicians and statisticians. One way to personalized medicine is to consider an individualized treatment strategy based on an individual’s characteristics that leads to the largest benefit. Recently, powerful machine learning methods have been proposed to estimate an optimal individualized treatment rule (ITR), but they are developed for binary treatment decisions and thus limited to compare only two treatments. When many treatment options are available, which is often the case in practice, existing methods need to be adapted by transforming a multicategory treatment selection problem into multiple binary treatment selections, for example, via one-vs-one or one-vs-all comparisons. However, how to combine multiple binary treatment selection rules into a single decision rule is not straightforward, and it is well known in the multicategory learning literature that some approaches may lead to inconsistent decision rules. In this article, we propose a novel and efficient method to generalize outcome weighted learning (O-learning) to multi-treatment settings. Specifically, we solve a multicategory treatment selection problem via sequential weighted support vector machines. Theoretically, we show that the resulting ITR is Fisher consistent. We demonstrate the performance of the proposed method with extensive simulations. An application to a three-arm randomized trial of treating major depressive disorder (MDD) shows that an individualized treatment strategy tailored to individual patients’ expectancy of treatment efficacy and their baseline depression severity reduces depressive symptoms more than non-personalized treatment strategies.

Fused Gaussian Process for Very Large Spatial Data
♦ Palong Ma and Emily Kang
University of Cincinnati
mapn@mail.uc.edu

With the development of new remote sensing technology, large or even massive spatial datasets covering the globe becomes available.
Statistical analysis of such data is challenging. This article proposes a semiparametric approach to model large or massive spatial datasets. In particular, a Gaussian process with additive components is proposed, with its covariance structure consisting of two components: one component is flexible without assuming a specific parametric covariance function but is able to achieve dimension reduction; the other is parametric and simultaneously induces sparsity. The inference algorithm for parameter estimation and spatial prediction is devised. The resulting spatial prediction method that we call fused Gaussian process (FGP), is applied to simulated data and a massive satellite dataset. The results demonstrate the computational and inferential benefits of the FGP over competing methods and show that the FGP is more flexible and robust against model misspecification.

Matrix Completion with Covariate Information

**Xiaojun Mao, Songxi Chen and Raymond Wong**
Iowa State University

This paper investigates the problem of matrix completion from corrupted data, when additional covariates are available. Despite being seldomly considered in the matrix completion literature, these covariates often provide valuable information for completing the unobserved entries of the high-dimensional target matrix \( A_0 \). Given a covariate matrix \( X \) with its rows representing the row covariates of \( A_0 \), we consider a column-space-decomposition model \( A_0 = X\beta_0 + B_0 \) where \( \beta_0 \) is a coefficient matrix and \( B_0 \) is a low-rank matrix orthogonal to \( X \) in terms of column space. This model decomposes \( A_0 \) orthogonally (in terms of column space) into two components allowing a clear separation between the interpretable covariate effects \( (X\beta_0) \) and the flexible hidden factor effects \( (B_0) \). Besides, our work allows the probabilities of observation to depend on the covariate matrix, and hence a missing-at-random mechanism is permitted. We propose a novel penalized estimator for \( A_0 \) by utilizing both Frobenius-norm and nuclear-norm regularizations, and achieve its practical computation with an efficient and scalable algorithm. Asymptotic convergence rates of the proposed estimators are studied. The empirical performance of the proposed methodology is illustrated via both numerical experiments and a real data application.

**Session 33: Multiple tests in Clinical trials**

**Bonferroni - based gatekeeping procedure with retesting option**

**Zhiying Qiu\(^1\), Wenge Guo\(^2\) and Sanat Sarkar\(^3\)**

\(^1\) Biostatistics and Programming, Sanofi US
\(^2\) New Jersey Institute of Technology
\(^3\) Temple University

In complex clinical trials, multiple research objectives are often grouped into sets of objectives based on their inherent hierarchical relationships. Consequently, the hypotheses formulated to address these objectives are grouped into ordered families of hypotheses and thus to be tested in a pre-defined sequence. In this paper, we introduce a novel Bonferroni based multiple testing procedure for testing hierarchically ordered families of hypotheses. The proposed procedure allows the families to be sequentially tested more than once with updated local critical values. It is proved to control the overall type I error rate strongly under arbitrary dependence. Implementation of the procedure is illustrated using two examples. Finally, the procedure is extended to testing multiple families of hypotheses with a complex two-layer hierarchical structure.

**A seamless phase II/III/IV clinical trial design with different endpoints for different phases**

**Hui Quan, Tianyue Zhou and Peng-Liang Zhao**
Sanofi

To save time and resources, a seamless phase II/III/IV clinical trial design is considered for a study. K doses compared to a placebo control are evaluated at phase II. Based on the results of a phase II endpoint, one of the K doses will be selected for phases III and IV. Patients of the selected dose and placebo control from phase II will stay until study completion and will be included in the phase III and IV analyses. Pre-planned additional number of patients will be enrolled into the selected dose and placebo control during phase III. Results of the phase III may be submitted for an early New Drug Application for an indication on the phase III endpoint. Pre-specified additional number of patients may be enrolled during phase IV. The phase IV endpoint is a time to event endpoint. Based on cumulative data up to phase III, adaptation on the total number of events for phase IV may be performed for desired power. After reaching the target number of events or the study completion, final analyses with all available data will be conducted for potential claims of treatment effects of the selected dose on the phase III and IV endpoints. A multiplicity adjustment procedure is proposed for the overall type I error rate control. Simulations are conducted to compare the performances of this design and other designs.
Session 34: Novel methods and practical considerations in missing data analysis

Tipping point sensitivity analysis for incomplete event count data

Xuezhou Mao1, Li Qi2, Hui Quan1, Mei Zhang1, Lin Wang1 and Lynn Wei1
1Biostatistics and Programming, Sanofi U.S.
2Translational Medicine, Sanofi U.S.

Over the recent years, the prevention and analysis of missing data have been deemed as one focus in clinical trials by the regulatory agencies, sponsors and academicians. Statistical sensitivity analysis for the missing mechanism assumption made by the primary analysis plays a critical role in examining the robustness or justifying the conclusions of the primary statistical model. This paper addresses the tipping point sensitivity analysis for the missing data occurred to the event count endpoint over the planned treatment period in a clinical trial, where premature treatment discontinuation and early study withdrawal lead potential event(s) not to be observed during the lost-to-follow-up period. We propose three methods that perform the sensitivity analysis based on different clinical interpretations, respectively. The first two methods are proposed based on parametric regression models and multiple imputations, while the other one does not rely on regression model and realizes the imputation on the missing cohort instead of at the subject level. We have applied the proposed methods in evaluating the robustness of the primary analysis to the incomplete count data of a real clinical trial. Practical considerations and recommendations are provided on applying each of the proposed methods under appropriate scenarios.

Options for implementing pattern-mixture-based sensitivity analyses

Ilya Lipkovich, Bohdana Ratitch and Michael O’Kelly
QuintilesIMS

Pattern-mixture-based modeling (PMM) strategies for clinical outcomes with missing values have attracted a lot of attention recently owing to their flexibility, clarity of assumptions, and interpretability of results for the end-users of statistical analysis: clinicians, regulators, and policy makers. A common use of PMMs was to provide sensitivity analyses to assess the robustness of a primary analysis that assumed outcomes were missing at random (MAR) to plausible departures from the MAR assumption, assuming the patients would have continued the assigned treatment. In addition, we note that the PMM can be used to evaluate an estimated that takes into account likely “real-world” outcomes under a variety of assumptions for patients who had discontinued from the assigned treatment for specific reasons. The flexibility of PMM stems from the fact that in most cases it can be implemented using the machinery of multiple imputation.

Diagnosing Missing Always at Random in Multivariate Data

Iavor Bojinov, Natesh Pillai and Donald Rubin
Harvard Statistics Department

Models for analyzing multivariate data sets with missing values require strong, often unassessable, assumptions. The most common of these is that the mechanism that created the missing data is ignorable - a twofold assumption dependent on the inferential procedure used. The first part, which is the focus here, under the Bayesian and direct likelihood paradigms, requires that the missing data are missing at random (MAR); in contrast, the frequentist-likelihood paradigm demands that the missing data mechanism always produces MAR data, a condition known as missing always at random (MAAR). Under mild conditions, assuming MAAR leads to an assumption that can be tested using the observed data alone namely, the missing data indicators only depend on the fully observed outcome variables. In this paper, we propose three different diagnostics procedures that not only indicate when this assumption is invalid but also suggest which variables are the most likely culprits. Although MAAR is not a necessary condition to ensure validity under the Bayesian and likelihood paradigms, it is sufficient, and evidence for its violation should encourage the statistician to carry out a targeted sensitivity analysis.

Multiple imputation of randomly censored covariates in regression analysis

Folefac Atem1, Jing Qian2, Jacqueline Maye3, Keith Johnson3 and Rebecca Betensky4
1University of Texas
2University of Massachusetts, Amherst
3Massachusetts General Hospital
4Harvard University

Randomly censored covariates arise frequently in epidemiologic studies. The most commonly used methods, including complete case and single imputation or substitution, suffer from inefficiency and bias, they make strong parametric assumptions and they consider limit of detection censoring only. We employ multiple imputation, in conjunction with semi-parametric modeling of the censored covariate, to overcome these shortcomings and to facilitate robust estimation. We develop a multiple imputation approach for randomly censored covariates within the framework of linear and logistic regression models. We use the nonparametric estimate of the covariate distribution, or the semi-parametric Cox model estimate in the presence of additional covariates in the model. We evaluate this procedure in simulations, and compare its operating characteristics to those from the complete case analysis and a survival regression approach. We apply the procedures to an Alzheimer’s study of the association between amyloid positivity and maternal age of onset of dementia.

Session 35: recent developments in parsimonious modelling

On the analysis of Bregman-surrogate algorithms for nonconvex optimization

Zhifeng Wang and Yiyuan She
Florida State University

Modern statistical problems often involve minimizing objective functions that are not necessarily convex or smooth. This paper proposes and investigates a broad surrogate framework defined by generalized Bregman divergence functions for developing scalable algorithms. Local linear approximation, mirror descent, iterative thresholding, and DC programming can all be viewed as particular algorithms. The Bregman re-characterization enables us to choose limit of detection censoring only. We employ multiple imputation, in conjunction with semi-parametric modeling of the censored covariate, to overcome these shortcomings and to facilitate robust estimation. We develop a multiple imputation approach for randomly censored covariates within the framework of linear and logistic regression models. We use the nonparametric estimate of the covariate distribution, or the semi-parametric Cox model estimate in the presence of additional covariates in the model. We evaluate this procedure in simulations, and compare its operating characteristics to those from the complete case analysis and a survival regression approach. We apply the procedures to an Alzheimer’s study of the association between amyloid positivity and maternal age of onset of dementia.

Rebecca Betensky
Folefac Atem
Jing Qian
Jacqueline Maye
Keith Johnson
Rebecca Betensky
University of Texas
University of Massachusetts, Amherst
Massachusetts General Hospital
Harvard University
Incorporating Covariates into Integrated Factor Analysis of Multi-View Data

Gen Li and *Sangkyu Jung

1 Columbia University
2 University of Pittsburgh
sangkyu@pitt.edu

In modern biomedical research, it is ubiquitous to have multiple data sets measured on the same set of samples from different views (i.e., multi-view data). For example, in genetic studies, multiple genomic data sets at different molecular levels or from different cell types are measured for a common set of individuals to investigate genetic regulation. Integration and reduction of multi-view data have the potential to leverage information in different data sets, and to reduce the magnitude and complexity of data for further statistical analysis and interpretation. In this paper, we develop a novel statistical model, called supervised integrated factor analysis (SIFA), for integrative dimension reduction of multi-view data while incorporating auxiliary covariates. The model decomposes data into joint and individual factors, capturing the joint variation across multiple data sets and the individual variation specific to each set respectively. Moreover, both joint and individual factors are partially informed by auxiliary covariates via nonparametric models. We devise a computationally efficient Expectation-Maximization (EM) algorithm to the model under some identical ability conditions. We apply the method to the Genotype-Tissue Expression (GTEx) data, and provide new insights into the variation decomposition of gene expression in multiple tissues. Extensive simulation studies and an additional application to a pediatric growth study demonstrate the advantage of the proposed method over competing methods.

Embracing the Blessing of Dimensionality in Factor Models

*Quefeng Li1, Guang Cheng2, Jianqing Fan3 and Yuan Wang3

1 UNC Chapel Hill
2 Purdue University
3 Princeton University quefeng@email.unc.edu

Factor modeling is an essential tool for exploring intrinsic dependence structures among high-dimensional random variables. Much progress has been made for estimating the covariance matrix from a high-dimensional factor model. However, the blessing of dimensionality has not yet been fully embraced in the literature: much of the available data is often ignored in constructing covariance matrix estimates. If our goal is to accurately estimate a covariance matrix of a set of targeted variables, shall we employ additional data, which are beyond the variables of interest, in the estimation? In this talk, we will provide sufficient conditions for an affirmative answer, and further quantify its gain in terms of Fisher information and convergence rate. In fact, even an oracle-like result (as if all the factors were known) can be achieved when a sufficiently large number of variables is used. The idea of utilizing data as much as possible brings computational challenges. A divide-and-conquer algorithm is thus proposed to alleviate the computational burden, and also shown not to sacrifice any statistical accuracy in comparison with a pooled analysis. Simulation studies further confirm our advocacy for the use of full data, and demonstrate the effectiveness of the above algorithm.

An efficient approach to discriminant analysis on tensor data with covariates

*Yuqing Pan, Qing Mai and Xin Zhang

Department of Statistics, Florida State University yuqinq.pan@stat.fsu.edu

In contemporary scientific research, it is of great interest to predict a categorical response with both scalar/vector covariates and high-dimensional tensor (i.e., multi-dimensional array) data. For example, brain imaging data usually contains a three-way tensor of the brain image and clinical measurements such as age, based on which we hope to predict the disease status. This mixture of different types of data leads to challenges in statistical analysis. On one hand, it is difficult to build interpretable models that explains the relationship between the covariates, the tensor and the response. On the other hand, these data often have intimidating dimensions such that the computation can be demanding even given the rich literature on high-dimensional data analysis. To tackle these challenges, we propose a hierarchical discriminant analysis model that integratively models the covariates and the tensor. Unlike many popular methods, we preserve the intrinsic structure of the data for interpretability. We further propose a penalized approach to obtain sparse estimates of the parameters. In order to fit this model, an efficient algorithm is implemented that takes advantage of the tensor structure. Theoretical results show that our method is consistent even when the dimension is much larger than the sample size. Both simulated and real data analysis supports the superior performance of our method over existing methods.

Session 36: Recent Advances in Enrichment Strategies: Methods and Case Studies

Bayesian Random Partition Models for Subgroup Identification

Wentian Guo1 and *Yuan Ji2

1 Laiya Consulting, Inc.
2 NorthShore U. HealthSys/The University of Chicago koaeraser@gmail.com

Personalized treatment strategies assign targeted therapies that are predicted to be effective based on the patient’s baseline characteristics such as biomarker profiles. As a result, subgroups of patients with similar profiles are formed and treated individually. Often times, patient subgroups are unknown but can be learned through inference using observed data. We present a class of Bayesian random partition models aiming to fulfill two simultaneous goals, 1) to enrich the allocation of each subgroup of patients to their precision and desirable treatments during a clinical trial and 2) to report multiple subgroup-treatment pairs (STPs) as an inference. Using random partitions and semiparametric Bayesian models, we provide coherent and probabilistic assessment of potential patient subgroups and their associated targeted therapies. Each STP can then be used for future confirmatory studies for regulatory approval. Multiple examples are provided ranging from trial designs and data analysis.

Optimization of Multi-arm Adaptive Enrichment Designs via Simulated Annealing

*Joshua Betz, Michael Rosenblum, Jon Steinigrsson and Adi Gherman

Johns Hopkins Bloomberg School of Public Health jbetz@jhu.edu

Adaptive enrichment designs allow for the pre-specified modification of enrollment criteria during an ongoing clinical trial based on accrued data, and have the potential to generate stronger evidence of benefit or harm in clinically relevant subpopulations. The operating characteristics of such trials, such as expected sample size...
On Group Sequential Enrichment Design for Basket Trial

**Shuai Yuan**, **Aiying Chen** and **Li He**

Cancer is becoming a collection of niche diseases defined by the molecular subtypes. Such understanding forms the basis of the basket trial, which pools together multiple cohorts of patients with the same molecular subtype across different tumor indications and thus facilitates the development of targeted therapies. Efficient pruning is critical in basket designs as it ensures the internal consistency of the selected indications and improves the probability of success on the selected pool of indications. In this article, we consider pruning both inactive and extremely active indications and propose a group sequential enrichment design with testing procedures to both control the family-wise error rate (FWER) in the weak sense and maintain the power for basket trials with such pruning strategy. We compare the proposed design with a relevant design which prunes only unresponsive subgroups. The total misclassification rates and misclassification rates among truly unresponsive indications are smaller for the proposed design, while the average power is similar.

**Case studies of Enrichment Strategies in Oncology**

**Ming Zhu**

Sanofi Pasteur

In FDA Guidance for Industry "Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products", enrichment is defined as the prospective use of any patient characteristic to select a study population in which detection of a drug effect (if one is in fact present) is more likely than it would be in an unselected population. Three broad categories of enrichment strategies are discussed in the Guidance: decreasing heterogeneity, prognostic enrichment and predictive enrichment. In this talk, a review of enrichment strategies will be presented and cases studies in oncology clinical trials will be discussed.

**Session 37: Recent robust statistical methodology in data analysis**

Robust likelihood inference for two correlated multinomial distributions

Tsung-Shan Tsou

Institute of Statistics, National Central University, Jhongli, Taiwan

We use the model for parallel designs to derive a robust score statistic for testing the equality of two multinomial distributions in paired designs. This test accounts for the within-cluster correlations in a data-driven manner without a full model specification that requires excessive nuisance joint probabilities. The robust score statistic reduces to the McNemar’s test in the paired binary data scenario. We use simulations and real data analysis to demonstrate the superiority of the robust procedure.

**Hypothesis testing in functional linear models**

Chongzhi Di, Yu-Ru Su and Li Hsu

Fred Hutchinson Cancer Research Center
cd@fredhutch.org

Functional linear models (FLM) are widely used to investigate the association between functional predictors and a scalar response variable. A popular approach to through dimension reduction by functional principal component analysis, however, inference is often sensitive to the choice of the number of principal components. We propose a new method of ordering and selecting principal components to construct test statistics. The proposed method takes into account both the association with the response and the variation along each eigenfunction. We establish its theoretical properties and assess the finite sample properties through simulations. Our simulation results show that the proposed test is more robust against the choice of threshold while being as powerful as, and often more powerful than, the existing method. We then apply the proposed method to the cerebral white matter tracts data obtained from a diffusion tensor imaging tractography study.

**Drug cumulative effect in observational study with application on ADHD patients.**

Shinsheung Yuan and Kang-chung Yang

Institute of Statistical Sciences, Academia Sinica

syuan@stat.sinica.edu.tw

Pharmacoepidemiology is the study using epidemiological methods to study long-term effects of drugs in population. In this study, we focused on the analysis of ADHD patients on their medications and the risks of other neurological diseases. The first question should be whether the cumulative effect exists or how big the time window is before applying the more advanced methods for estimating the weights. We proposed to use the simple bootstrap approach for choosing the variables and determining the window size. Moreover, the complicated medication history will also influence the estimates of the weights and the correct selection of patients is needed for reducing such biases. Finally, discussion and results of our analyses will be provided.

**Session 38: Modern Developments in Statistical Classification and Selection**

Model-free variable selection for the regression mean

Yuexiao Dong

Temple University

ydong@temple.edu

A novel test statistic is proposed to identify important predictors for the conditional mean function in regression. The stepwise regression algorithm based on the proposed test statistic guarantees variable selection consistency without specifying the functional form of the conditional mean. When the predictors are ultrahigh dimensional, a model-free screening procedure is introduced to precede the stepwise regression algorithm. The screening procedure has the sure screening property when the number of predictors grows at an exponential rate of the available sample size. The finite-sample performances of our proposals are demonstrated via numerical studies.
High-dimensional interaction selection via regularization

Ning Hao, Yang Feng and Helen Zhang

1University of Arizona
2Columbia University
yang.feng@columbia.edu

Quadratic regression (QR) models naturally extend linear models by considering interaction effects between the covariates. To conduct model selection in QR, it is important to maintain the hierarchical model structure between main effects and interaction effects. Existing regularization methods generally achieve this goal by solving complex optimization problems, which usually demands high computational cost and hence are not feasible for high-dimensional data. This paper focuses on scalable regularization methods for model selection in high-dimensional QR. We first consider two-stage regularization methods and establish theoretical properties of the two-stage LASSO. Then, a new regularization method, called Regularization Algorithm under Marginality Principle (RAMP), is proposed to compute a hierarchy-preserving regularization solution path efficiently. Both methods are further extended to solve generalized QR models. Numerical results are also shown to demonstrate the performance of the methods.

An Umbrella Algorithm for Neyman-Pearson Classification

Xin Tong, Yang Feng and Jingyi Li

1University of Southern California
2Columbia University
xint@marshall.usc.edu

In many binary classification applications, such as disease diagnosis and spam detection, practitioners commonly face the need to limit type I error (i.e., the conditional probability of misclassifying a ‘normal’, or class 0, observation as ‘abnormal’, or class 1) so that it remains below a desired threshold. To address this need, the Neyman-Pearson (NP) classification paradigm is a natural choice; it minimizes type II error (i.e., the conditional probability of misclassifying a class 1 observation as class 0) while enforcing an upper bound, α, on the type I error. Although the NP paradigm has a century-long history in hypothesis testing, it has not been well recognized and implemented in statistical classification schemes. Common practices that directly limit the empirical type I error to no more than α do not satisfy the type I error control objective because the resulting classifiers are still likely to have type I errors much larger than α. As a result, the NP paradigm has not been properly implemented for many classification scenarios in practice. In this work, we develop the first umbrella algorithm that implements the NP paradigm for all scoring-type classification methods, including popular methods such as logistic regression, support vector machines and random forests.

Model-based Clustering for Dynamic Networks

Kevin Lee, Lingzhou Xue and David Hunter

The Pennsylvania State University
kh1198@psu.edu

Dynamic network modeling provides an emerging statistical technique for various real-world applications. It is a fundamental research question to detect the community structure in dynamic networks. However, due to significant computational challenges and difficulties in modeling communities, there is little progress in the current literature to effectively find communities in dynamic networks. In this work, we propose a novel model-based clustering framework for dynamic networks, which is based on exponential-family random graph models and inherits the philosophy of finite mixture models. To determine an appropriate number of communities, a conditional likelihood Bayesian information criterion is proposed. Moreover, an efficient variational expectation-maximization algorithm is designed to solve approximate maximum likelihood estimates of network parameters and mixing proportions. By using variational methods and minorization-maximization techniques, our methods have appealing scalability for large scale dynamic networks. Finally, the power of our methods is demonstrated in real applications to dynamic international trade networks.

Session 39: Non- and semi-parametric models in the large-scale genomic data analyses

Improved shrunken centroid classification for better variable selection in genomics data

Xukan Li, Haiyan Wang and Huaiyu Zhang

Kansas State University
hwang@ksu.edu

In this talk, I will present an extension of the popular nearest shrunken centroid classification (PAM) in high dimensional data. PAM performs variable selection via soft-thresholding the statistics formed by comparing class centroids with the overall centroid. A common thresholding parameter is selected via cross validation and is used on all variables. This imposes the constraint that the statistics from all variables and all classes need to be identically distributed. This is typically violated in genome studies since the data from high throughput equipment are often fluorescence measurements of light intensity which are heterogeneous and highly skewed. Consequently, PAM has high false positive rates. Here we introduce a scheme to provide weighted thresholding for variable selection. A unique weight is calculated for each comparison statistic so that different variables are allowed to have different distributions. Computation of the weight is based on Edgeworth expansion and Cornish-Fisher expansion of the comparison statistic. The resulting algorithm significantly reduces false positive rates and improves classification accuracy in simulation and real data applications.

A multivariate semiparametric approach for gene-environment interactions

Cen Wu, Yuehua Cui and Shaungge Ma

1Department of Statistics, Kansas State University
2Dept. of Stat. and Prob. Michigan State University
3Department of Biostatistics, Yale University
wucen@ksu.edu

Gene-environment interactions play important roles in elucidating the genetic basis of complex disease. It has been traditionally pursued by assuming linear relationship between genetic and environment effects, along with their interactions. To accommodate possible nonlinear effects of some environmental exposure, we propose a multivariate semi-parametric model and adopt penalization to select important markers. The proposed method can flexibly identify structural gene environment interactions, and can be efficiently implemented using a coordinate descent algorithm. Both simulation studies and a case study on the Health Professionals Follow-up Study demonstrate the advantage of the proposed method over alternatives.

Bayesian integrative model to understand joint genetic epigenetic effects in development of eczema

Yu Jiang, Cen Wu, Hongmei Zhang, Su Chen and Wilfried Karmass

100

2017 ICSA Applied Statistics Symposium, Chicago, June 25-28
Eczema, also called atopic dermatitis, is one of the most common cutaneous disease that occurs as early as in infancy. It is believed that genetic susceptibility, environmental factors, and potential interactions between genes, plays crucial roles in disease development. Identifying potential molecular markers are critical next steps in the development of effective strategies for disease prevention and treatment. Most existing genetic and epigenetic disease association studies focus on one single type of profiling data. Analyzing a single type of omics measurement cannot comprehensively describe the underlying biological process and will result in false or non-informative molecular markers. More recent studies moved toward integrating different types of measures to assess the joint contribution of genetic and epigenetic factors. In the current study, we develop a Bayesian semi-parametric integrative approach which integrates both genetic and epigenetic data and have the ability to identify 1) SNPs and CpGs that track with or influence a certain gene’s expression and thus affect the disease outcomes and 2) genes associated with disease phenotype by taking into account the regulatory effects of SNPs and DNAm. The methods are evaluated by extensive simulations and applied to multidimensional omics datasets from the Isle of Wight (IOW) third generation cohort.

An integrative method for joint analysis of multi-level omic data
Pei Geng, Xiaoran Tong and Qing Lu
Michigan State University
gengpe1@stt.msu.edu

With rapidly evolving high-throughput technologies and ever-decreasing costs, it has become feasible to systematically study diverse types of omic data in biological and clinical studies. We develop a nonparametric integrative U method for the design and analysis of multi-level omic studies. More specifically, targeting on the joint analysis of biomarker information and gene expression data, we applied the functional data framework on the genetic variants then proposed the test based on U statistic. The proposed method offers several attractive advantages: first, it is free of model assumptions; second, it can accommodate different types of phenotypes such as Gaussian, binary even heavy-tailed continuous distributions; third, it is computationally efficient without complicated approximation procedures due to the asymptotic normality. Simulation results showed stable empirical levels and superior power performance compared to variance component tests under various scenarios. The method was also applied to the San Antonio Family Heart Study and the San Antonio Family Diabetes/Gallbladder Study to detect significant genes.

Session 40: Integrating metagenomics and other -omics data

Mediation Analysis in Microbiome Studies
Hongze Li
University of Pennsylvania
hongze@upenn.edu

Motivated by advances in the causal inference on mediation and problems arising in the analysis of metagenomic data, we consider the effect of a treatment on an outcome transmitted through compositional mediators. Compositional and high dimensional features of such mediators make the standard mediation analysis not directly applicable. A sparse mediation model for high-dimensional compositional data is proposed in this paper utilizing the algebraic structure of a composition under the simplex space and a constraint linear regression model to achieve subcompositional coherence. Under this model, we develop estimation method for estimating indirect microbial mediation effect and direct effect of a randomly assigned treatment on the outcome and their variances using bootstrap. Tests of overall mediation effect of all bacterial taxa and individual mediation effect are also proposed. We conduct extensive simulation studies to assess the performance of the proposed method and apply the method to a real metagenomic dataset to investigate the effect of fat intake on body mass index (BMI) mediated through the gut microbiome composition.

An Optimum Microbial Association Test
Hyunwook Koh, Martin Blaser and Huilin Li
New York University School of Medicine
huilin.li@nyumc.org

Background: The role of the microbiota in human health and disease has been increasingly studied, gathering momentum through the use of high-throughput technologies. Further identification of the roles of specific microbes is necessary to better understand the mechanisms involved in diseases related to microbiome perturbations. Here, we introduce a new microbiome-based group association testing method, optimal microbiome-based association test (OMiAT). OMiAT is a data-driven testing method which takes an optimal test throughout different tests from the sum of powered tests (SPU) and microbiome regression-based kernel association test (MiRKAT). We illustrate that OMiAT efficiently discovers significant association signals arising from varying microbial abundances and different relative contributions from microbial abundance and phylogenetic information. We also propose a way to apply it to fine-mapping of diverse upper-level taxa at different taxonomic ranks (e.g., phylum, class, order, family, and genus), as well as the entire microbial community, within a newly introduced microbial taxa discovery framework, microbiome comprehensive association mapping (MiCAM). Results: Our extensive simulations demonstrate that OMiAT is highly robust and powerful compared with other existing methods, while correctly controlling type I error rates. Our real data analyses also confirm that MiCAM is especially efficient for the assessment of upper-level taxa by integrating OMiAT as a group analytic method. Conclusions: OMiAT is attractive in practice due to the high complexity of microbiome data and the unknown true nature of the state. MiCAM also provides a hierarchical association map for numerous microbial taxa and can also be used as a guideline for further investigation on the roles of discovered taxa in human health and disease.

Integrative Analysis of High-dimensional Microbiome and Host Genome Data
Lingling An and Meng Lu
University of Arizona
anling@email.arizona.edu

Microbiome can have a major impact on the phenotype of their host, e.g., human health. Recent researches have been focused on association between human microbiota and the related diseases. In this research we are interested in understanding how the microbiome and host genome jointly impact human health and disease by integrating multiple “omics” technologies through mediation analysis. Human genome is treated as mediator and microbiome as exposure. Currently there is lack of analytic approaches to dealing with high-dimensional correlated mediators (e.g., some genes are correlated due to a biological pathway) without any data transformation. We
propose a two-stage regression approach to identifying the association between microbiome, human genome, and human disease status under a high dimensional setting. Joint significance test is used to evaluate the significance of the multiple correlated mediators while controlling the family-wise error rate.

Session 41: Design and Uncertainty Quantification for Computer Experiments

A Bayesian Approach to Model Selection & Uncertainty Quantification in Brain Injury Simulations

Sandeep Madireddy, Kamar Vemaganti and Emily Kang
Argonne National Laboratory

Computational models of the head and brain are extensively used to study traumatic brain injuries (TBI). Simulations of TBI are complex and computationally challenging because of (a) the large uncertainty in the material testing data and the resulting uncertainty in the constitutive model parameters, (b) high strain rates and the short time scales of the impact loading, and (c) the complex geometry of the human brain model. We propose a Bayesian framework for parameter estimation and constitutive model selection based on the parallel nested Monte Carlo sampling algorithm MULTINEST that can be used to address some of these challenges. Four different factors are used to reliably choose a parsimonious model from the candidate set of models. These are the qualitative fit of the model to the experimental data, evidence values, maximum likelihood values, and the landscape of the likelihood function. This approach provides a robust and efficient alternative to Markov chain Monte Carlo methods to sample from multi-modal distributions and to efficiently calculate the evidence integral.

We model the brain tissue as a non-linear visco-hyperelastic material and used the proposed framework for simultaneous model selection and estimation of the posterior distributions of the parameters. This material model is implemented into the commercial finite element solver LS-DYNA and the traumatic brain injury caused by impact loading from a vehicle crash is simulated using the SIMon finite element model of the human head. A maximum principal strain-based injury criterion is used to assess the severity of the brain injury. The uncertainty in the material model parameters is non-intrusively propagated to the injury criterion using a Bayesian Gaussian process surrogate of the finite element model to avoid computationally expensive simulations. This enables us to calculate the probability that an injury tolerance is reached for a given impact loading. This probabilistic method can be used to further simulate injuries and validate various injury criteria with high fidelity.

Learning About Physical Parameters - The Importance of Model Discrepancy

Jenny Brynjarsdottir and Anthony OHagan
1 Case Western Reserve University
2 University of Sheffield

Science-based simulation models are widely used to predict the behavior of complex physical systems. It is also common to use observations of the physical system to solve the inverse problem, that is, to learn about the values of parameters within the model, a process which is often called calibration. The main goal of calibration is usually to improve the predictive performance of the simulator but the values of the parameters in the model may also be of intrinsic scientific interest in their own right. In order to make appropriate use of observations of the physical system it is important to recognize model discrepancy, the difference between reality and the simulator output. We illustrate through a simple example that an analysis that does not account for model discrepancy may lead to biased and over-confident parameter estimates and predictions. The challenge with incorporating model discrepancy in statistical inverse problems is being confounded with calibration parameters, which will only be resolved with meaningful priors. For our simple example, we model the model-discrepancy via a Gaussian process and demonstrate that through accounting for model discrepancy our prediction within the range of data is correct. However, only with realistic priors on the model discrepancy do we uncover the true parameter values. Through theoretical arguments we show that these findings are typical of the general problem of learning about physical parameters and the underlying physical system using science-based mechanistic models.

A General Construction for Space-filling Latin Hypercubes

Chunfang Lin and Lulu Kang
1 Queen’s University, Canada
2 Illinois Institute of Technology

We propose a general method for constructing Latin hypercubes of flexible run sizes for computer experiments. The method makes use of arrays with a special structure and Latin hypercubes. By using different such arrays and Latin hypercubes, the proposed method produces various types of Latin hypercubes including orthogonal and nearly orthogonal Latin hypercubes, sliced Latin hypercubes, and Latin hypercubes in marginally coupled designs. In addition, the proposed algebraic design construction is particularly efficient as it does not need any optimization search but still, produces Latin hypercubes whose space-filling properties are comparable with those generated by the common and latest methods in the literature.

Sequential design of experiment for multiple computer models

Bledar Konomi
University of Cincinnati

For a physical system there are usually a number of different computer models developed with different underlying physical mechanism, level of detail, resolution and tuning parameters. In spite of the availability of powerful computational resources, it often is computationally too expensive to run such complex models for all possible input and boundary conditions. Separate and independent analysis may result in weak statistical inference because information from different sources with different details is ignored. In the UQ literature there are a number of methods based on the autoregressive co-kriging mode to build emulator models. One of the key issue in an emulator model is to optimally choose the design points for computer simulator based on which the emulator is built. Methods, such as Latin hypercube sampling (LHS) and space filing designs, orthogonal arrays, and multilevel Monte Carlo (MC), have been developed to sample the input space without considering information about the output. More sophisticated methods use the active learning sequential design of experiment in the case of one computer model. In this work we extend active learning sequential design methods to multiple computer models.
Session 42: Statistical Methods for Network Data

A new SVD approach to optimal topic estimation
• Tracy Ke and Minche Wang
University of Chicago
zke@galton.uchicago.edu
In the probabilistic topic models, the quantity of interest’s a low-rank matrix consisting of topic vectors is hidden in the text corpus matrix, masked by noise, and Singular Value Decomposition (SVD) is a potentially useful tool for learning such a low-rank matrix. However, the connection between this low-rank matrix and the singular vectors of the text corpus matrix are usually complicated and hard to spell out, so how to use SVD for learning topic models faces challenges.
We overcome the challenge by revealing a surprising insight: there is a low-dimensional simplex structure which can be viewed as a bridge between the low-rank matrix of interest and the SVD of the text corpus matrix, and which allows us to conveniently reconstruct the former using the latter. Such an insight motivates a new SVD-based approach to learning topic models.
For asymptotic analysis, we show that under the popular probabilistic topic model (Hofmann, 1999), the convergence rate of the 11-error of our method matches that of the minimax lower bound, up to a multi-logarithmic term. In showing these results, we have derived new element-wise bounds on the singular vectors and several large-deviation bounds for weakly dependent multinomial data. Our results on the convergence rate and asymptotical minimaxity are new. We have applied our method to two data sets, Associated Process (AP) and Statistics Literature Abstract (SLA), with encouraging results. In particular, there is a clear simplex structure associated with the SVD of the data matrices, which largely validates our discovery.

Regression analysis of networked data
Yan Zhou1 and • Peter Song2
1Merck & Co.
2University of Michigan
pxsong@umich.edu
We develop a new regression analysis approach to evaluating associations of covariates with outcomes measured from networks. This development is motivated from a study of infant growth that collects outcomes of event related potentials (ERP, a type of neuroimaging) measured over electroencephalogram (EEG) electrodes on the scalp. We propose a new generalized method of moments (GMM) that incorporates both established and data-driven knowledge of network topology among nodes in the estimation and inference to achieve robustness and efficiency. The GMM approach is computationally fast and stable to handle the regression analysis of network data, and conceptually it is simple with desirable properties in both estimation and inference. Both simulation studies and real EEG data analysis will be presented for illustration.

Time-varying network estimation from high-dimensional time series
• Mengyu Xu1, Xiaohui Chen2 and Wei Biao Wu3
1University of Central Florida
2University of Illinois Urbana-Champaign
3University of Chicago
Mengyu.Xu@ucf.edu
The estimation of time-varying networks from high-dimensional time series is considered. Two types of non-stationarity are investigated: structural breaks and smooth changes. Our approach can achieve consistent detection of the change points and simultaneous estimation of the piece-wise smoothly-varying networks. Rates of convergence for estimating change points and networks are obtained under mild moment and dependence conditions. The method is applied to the analysis of network structure of the S&P 500 index.

Testing and Estimation of Social Network Dependence with Time to Event Data
Rui Song
North Carolina State University
rsong@ncsu.edu
Nowadays, events are spread rapidly along social network since we can share information with friends easily. We are interested in how people are affected by their friends’ behavior. For example, if a person share the game he or she is playing, will his or her friends start playing it as well? Studying social network dependence is an emerging research area. In this work, we propose a novel latent spatial auto-correlation Cox model to study social network dependence with time-to-event data. The proposed model introduces a latent indicator to characterize whether a person might be affected by his or her friends’ behavior. We first propose a score-type test for detecting the existence of social network dependence. If it exists, we further develop an EM-type algorithm to estimate the model parameters. The performance of the proposed test and estimators are illustrated by simulation studies and an application to a time-to-event data set about playing a popular QQ game from Tencent.

Session 43: Selected Topics in Designing and Conducting Safety Studies

Post-market evaluation of drug safety
John Yap1, Brenda Crowe2, • Simione Pinheiro3 and Yi Yvonne Huang3
1Food and Drug Administration
2Eli Lilly and Company
3University of Maryland
simone.pinheiro@fda.hhs.gov
This presentation will discuss the role of population-based assessments in drug safety evaluation and regulation. In this presentation, I will discuss the main strengths and limitations of evidence sources available for evaluation of post-market drug safety with particular focus on the contribution of population-based assessments in the evaluation of the safety of approved drugs and therapeutic biologies. These include studies available in the published medical literature, FDA-conducted assessments including the Sentinel active surveillance system, and post-market requirements. This presentation will also highlight the importance of the totality of the evidence in informing regulatory decision making.

Safety Signal Detection when Incidence Percentages are Biased
Brenda Crowe
Eli Lilly and Company
crowe_brenda.j@lilly.com
When comparing treatment emergent adverse events between treatment groups in a fixed duration controlled treatment period (i.e., the protocol plans for same follow-up duration in each arm), incidence percentages, together with a metric such as risk difference, risk ratio or odds ratio are often sufficient for safety signal detection. For combined analysis of multiple studies, study-size-adjusted percentages are a natural analogue. However, in some cases incidence percentages may give biased treatment effect results. In such cases, other metrics, such as exposure-adjusted incidence rates or time-to-event, may be appropriate. Of course, each of these can give biased results in some situations, so careful thought should be given to the
Meta-analysis of observational studies for drug safety

John Yap

FDA

There are many published observational studies on drug safety available in the scientific literature. Researchers often find value in these types of studies because, when compared to randomized clinical trials (RCTs), they reflect reality more closely in the sense that patients are followed and observed based on the use of drugs in practice. When a marketed drug shows a safety concern for an adverse outcome, observational studies may therefore provide useful information. In the case when multiple observational studies of a particular drug or treatment are available, one is often faced with the issue of how to summarize the study findings. A more commonly used approach over recent years is to perform a meta-analysis in order to obtain a summary numerical value of the risk of an outcome. While the use of meta-analysis in RCT’s has been extensively discussed in books and other sources, its use in observational studies has not been addressed adequately. In this presentation, we discuss the challenges and limitations when performing a meta-analysis of observational studies.

Latent Propensity Score Approach Allowing Covariate Measurement Error in Observational Studies

Yi Huang1, Elande Baro2, Andrew Raim3, Anindya Roy3 and Karen Bandeen-Roche4

1Dept. of Math & Stat, UMBC
2OB/OTS/CDER, FDA
3US Census Bureau
4Dept. of Biostatistics, Johns Hopkins University.

For observational studies, we often run into covariates which are measured with unobserved errors, particularly for medical, epidemiological, and health policy research. Unfortunately, this measurement error will thread the validity of causal inference based on the standard propensity score framework (Rosenbaum, Rubin, 1983, 1984) given the violation of its nonconfoundedness assumption. If we ignore such error (i.e. pretending all observed covariates are error free) and use the popular propensity score approach for treatment effect estimation, it would lead to biased results as we showed in paper. With the extend causal framework allowing covariate measurement error, we provided a latent propensity score approach accounting for such errors by identifying the homogeneous subgroups of propensity scores stochastically rather than deterministically. We developed both EM algorithm and Bayesian estimation methods for the proposed joint likelihoods combining outcome models and propensity score models in a finite mixture modeling framework. Extensive simulation studies are discussed to show the numerical performance of this newly developed approach by various estimation methods, and compared to the standard propensity score approach ignoring such errors. This is a joint work with my colleagues and students.

Session 44: Advanced Bayesian Methods with Applications to Medical Data

Bayesian semiparametric analysis of mixed effects models with applications to dose-response studies

Taeryon Choi

Korea University

In this talk, we present fully Bayesian approaches to semiparametric additive mixed effects models for analyzing either longitudinal data or clustered data with applications to dose-response studies. In the semiparametric mixed effects model structure, we estimate nonparametric smoothing functions of continuous covariates by using a spectral representation of Gaussian processes and the subject-specific random effects by using Dirichlet process mixtures. In this framework, we develop semiparametric mixed effects models that include normal regression and quantile regressions with or without shape restrictions. In addition, we deal with the Bayesian nonparametric measurement error models, or errors-in-variable regression models, using Fourier series and Dirichlet process mixtures, in which the true covariate is not observable, but the surrogate of the true covariate, is only observed. The proposed methodology is compared with other existing approaches to additive mixed models in simulation studies and benchmark data examples. More importantly, we consider two real data applications for dose-response analysis, in which measurement errors and shape constraints in the regression functions need to be incorporated with inter-study variability.

Flexible Priors for Covariance Functions

Dennis Cox

Rice University

dcox@rice.edu

We explore properties of inverse Wishart processes which model random covariance functions and may be used as priors for nonparametric smoothing and functional data. These are very convenient since all finite dimensional marginals are inverted Wishart distributions. Thus, they form a convenient class of priors for Bayesian modeling. We investigate properties of the process inherited from the scale operator and degrees of freedom.

Bayesian region selection in functional regression

Hongxiao Zhu and Yizhi Sun

Virginia Tech

hongxiao@vt.edu

Detecting important regions in functional data has great significance in guiding decision-making. Although variable selection among multiple functional predictors is by no means new, the selection of unknown regions within a functional object has been less studied. We propose a novel Bayesian method for region selection in the framework of functional data regression. The selection of regions is achieved through encouraging sparse estimation of the regression coefficient. Nonzero regions of the estimated coefficient function correspond to the selected regions. In particular, we adopt compactly supported and potentially over-complete basis to capture local features of the regression coefficient function, and assume spike-slab priors to coefficients of the bases functions. To encourage continuous shrinkage of nearby regions, we adopt an Ising hyper-prior to take into account the neighboring structure of the bases functions, represented by an undirected graph. Posterior sampling is performed through Markov chain Monte Carlo algorithms. We finally verify the practical performance by applying the proposed approach to the near-infrared and sonar data.
Session 45: Statistical and Regulatory Considerations and Opportunities for Rare Diseases Clinical Trials

Novel global statistical tests for multiple outcomes in clinical trials
♦ Michael McDermott and Donald Hebert
1 University of Rochester Medical Center
2 EMMES Corporation
Michael.McDermott@urmc.rochester.edu

Multiple primary outcome variables are sometimes used in clinical trials when it is important to capture different domains of the disease, all of which have the potential to be affected by treatment. A global statistical test is one strategy for analyzing multiple outcomes that can greatly enhance power to detect treatment effects, a particularly important consideration in trials involving rare diseases. Commonly used tests have excellent power when the treatment effects are consistent across all outcomes but can have very poor power against more general alternatives. We propose and study a novel class of global statistical tests based on procedures for combining p-values that retain good power properties for a wide variety of alternatives of clinical interest. Extensions to outcomes of mixed type are discussed. The tests are motivated by recently completed and ongoing clinical trials in neurological disease.

Regulatory Considerations and Opportunities for Rare Disease Clinical Trials
John Scott
FDA / CBER
john.scott@fda.hhs.gov

Developing therapeutic products for very rare diseases presents enormous challenges in clinical trial design and analysis. In order for a drug or biologic product to be approved in the United States, it needs first to be shown to be safe and effective. In this talk, I will consider some Bayesian and adaptive approaches that could be considered for assessing the effectiveness and safety of products for small populations. I will also present case examples of rare disease products reviewed by the Center for Biologics Evaluation & Research at the U.S. Food and Drug Administration, focusing on design features that facilitated determinations of safety and effectiveness based on very small clinical trials, sometimes involving as few as a dozen subjects.

Adult Data Extrapolation Models for Improved Efficacy and Toxicity Estimation in Pediatric Trials
♦ Cynthia Basu and Bradley P. Carlin
School of Public Health, University of Minnesota.
basu044@umn.edu

In most regulatory settings, the usual Phase I-II-III approval process is accompanied by extensive studies performed to understand the drug’s pharmacokinetic (PK) and pharmacodynamic (PD) properties. Although the results of these early stages are used in scale-up to derive first-in-man drug doses, these detailed data are not directly used in the clinical testing stage. In this talk, we attempt to utilize the rich PK/PD data obtained in earlier clinical work to inform the borrowing of information from adults when studying pediatric patients during the later clinical phases of drug development. In pediatric settings, it is especially crucial that we are parsimonious with the patients recruited for experimentation. We will use population PK/PD modeling to quantitatively assess the similarity between adults and children, and use this information in various hierarchical Bayesian adult borrowing rules whose statistical properties can then be evaluated. In particular, we will discuss methods to simulate the bias and mean square error to assess the performance of our approaches in settings where borrowing is and is not warranted to inform guidelines for the future use of our methods.

Session 46: Adaptive and Innovative Study Designs in Clinical Trials

Biomarker-driven clinical trial design in precision medicine
Jing Wang
Pfizer, Inc.
jing.wang122@pfizer.com

Precision medicine has paved the way for a new era of delivering tailored treatment options to patients according to their biological profiles. In combination with innovative adaptive design, this has presented drug developers unprecedented opportunities to engage novel thinking to accelerate drug discovery. This presentation will cover both classical and adaptive designs with biomarkers. Design options for biomarkers with very strong credentials, strong credentials and weak credentials will be discussed. Related statistical theories and analysis strategies will also be presented with case studies.

New CARA Designs for Personalized Medicine and Their Statistical Inference
♦ Wanying Zhao and Feifang Hu
The George Washington University
wzhao14@gwu.edu

Precision medicine takes account of individual characteristics for disease treatment and prevention. To develop precision medicine, more covariates of patients are under consideration in clinical trials. Based on different roles played in clinical studies, covariates can be categorized into two types: (1) prognostic covariates, which are used to balance treatment allocations for a simple treatment comparison, and (2) predictive covariates, which are used for treatment selection based on efficiency and ethics. We propose a general framework of new designs, which can incorporate both prognostic and predictive covariates in the randomization procedure simultaneously. Theoretical properties of statistical inference under new designs are provided based on linear models when the prognostic covariates, which are balanced in randomization, are excluded from the working model of inference. Some mild conditions for imbalances and target allocation proportions need to be satisfied to derive asymptotic properties of test statistics. It is proved that the test for comparing treatment effects is usually conservative in terms of small Type I error. One possible solution to this problem is the bootstrap method. New designs have advantages on improving average outcomes, but still allowing statistical inference with high power.

Seamless Phase 2/3 Study Design with an Oncology Example
♦ Zhaoyang Teng, Guohui Liu and Yi Liu
Takeda Pharmaceuticals
zhaoyang.teng2@takeda.com

Seamless phase 2/3 clinical trials are conducted in two stages with treatment selection or Go/No-Go decision at the first stage and efficacy confirmation at the second stage. Seamless phase 2/3 adaptive trial has a few advantages compared to the traditional phase 2 followed by phase 3 approach. First, it reduces the lead time between studies. In practice, the lead time between phase 2 study and phase 3 study is about 6-12 months. Second, the nature of adaptive design will also allow the investigator to make Go/No-Go decision at the end of stage 1. Third, seamless phase 2/3 may require a smaller sample size compared to the traditional approach which conduct phase 2 and phase 3 separately. In this research, we will discuss
how to determine the optimal sample size needed to make Go/No-Go decision at phase 2 portion by controlling both false positive rate and false negative rate. The detailed algorithm of determination of Go/No-Go criteria will also be provided. Moreover, we will discuss how to add an interim analysis at phase 2 portion to speed up the Go/No-Go decision making. In the end, a real oncology example will be used to illustrate the proposed seamless phase 2/3 study design.

Optimal flexible sample size design with interim dose determination

• Lu Cai and Lanju Zhang
AbbVie Inc.
lu.cui@abbvie.com

Confirmatory phase III clinical trials to evaluate the therapeutic effect of a new drug may have multiple doses involved. Traditionally, evaluation is performed at the end of the study and the primary efficacy analysis is conducted with multiplicity adjustment to enforce type I error rate control. Using interim analysis, the trial sponsor may want to stop the study early for futility if no dose looks promising or terminate a few less performing doses while letting others continue to the end. The present research considers such a trial with interim dose determination based on continuous treatment outcome. To achieve robust power, sample size is re-estimated and the design is optimized with respect to various design parameters. The presentation includes suggested interim dose determination criterion and final analysis procedure with relevant formulae adjusted for multiplicity involved. Simulation results follow as illustration.

Session 47: New machine learning tools for complex data

Sparse Tensor Response Regression

• Will Wei Sun1 and Lexin Li2
1University of Miami
2University of California, Berkeley
wsun@bus.miami.edu

Motivated by applications in Neuroimaging analysis, we propose a new regression model with a tensor response and a vector predictor. The model embeds two key sparse structures: element-wise sparsity and low-rankness. It can handle both a general and a symmetric tensor response, and thus is applicable to both structural and functional Neuroimaging data. We formulate the model parameter estimation as a non-convex optimization problem, and develop an efficient alternating updating algorithm. We establish a non-asymptotic estimation error bound for the actual estimator obtained from the proposed algorithm. This error bound reveals an interesting interaction between the computational efficiency and the statistical rate of convergence. Based on this general error bound, we further obtain an optimal estimation error rate when the distribution of the error tensor is Gaussian. We illustrate the efficacy of our model through intensive simulations and an analysis of the Autism spectrum disorder Neuroimaging data.

GENERAL FRAMEWORK FOR ASSOCIATION ANALYSIS OF HETEROGENEOUS DATA

• Gen Li3 and Irina Gaynanova2
1Columbia University, Department of Biostatistics
2Texas A&M University, Department of Statistics
guanyu@buffalo.edu

Multivariate association analysis is of primary interest in many applications. Despite the prevalence of high-dimensional and non-Gaussian data (such as count-valued or binary), most existing methods only apply to low-dimensional datasets with continuous measurements. Motivated by the Computer Audition Lab 500-song (CAL500) music annotation study, we develop a new framework for the association analysis of two sets of high-dimensional and heterogeneous (continuous/binary/count) data. We model heterogeneous random variables using exponential family distributions, and exploit a structured decomposition of the underlying natural parameter matrices to identify shared and individual patterns for two datasets. We also introduce a new measure of the strength of association, and a permutation-based procedure to test its significance. An alternating iteratively reweighted least squares algorithm is devised for model fitting, and several variants are developed to expedite computation and achieve variable selection. The application to the CAL500 data sheds light on the relationship between acoustic features and semantic annotations, and provides an effective means for automatic annotation and music retrieval.

Hierarchical models for predicting above ground biomass with 3D LiDAR signals

• Yuzhen Zhou1, Andrew Finley2 and Bruce Cook3
1University of Nebraska Lincoln
2Michigan State University
3NASA
yuzhenzhou@unl.edu

Recent advancements in remote sensing technology, specifically Light Detection and Ranging (LiDAR) sensors, provide the data needed to quantify forest characteristics at a fine spatial resolution over large geographic domains. We proposed a process-based Bayesian hierarchical model for the prediction of above ground biomass (AGB) using the 3D LiDAR signals. We offer simulation experiments to evaluate our proposed models and also apply them to a real dataset comprising LiDAR and spatially coinciding forest inventory variables.

Optimal Sparse Linear Prediction for Block-missing Multimodality Data without Imputation

• Guan Yu1, Quefeng Li2, Dinggang Shen2 and Yufeng Liu2
1State University of New York at Buffalo
2University of North Carolina at Chapel Hill
guanyu@buffalo.edu

In modern scientific research, data are often collected from multiple modalities. Since different modalities could provide complementary information, statistical prediction methods using multi-modality data could deliver better prediction performance than using single modality data. However, one special challenge for using multi-modality data is related to block-missing data. In practice, due to dropouts or the high cost of measures, the observations of a certain modality can be missing completely for some subjects. In this paper, we propose a new Direct Sparse regression procedure using COvariance from Multi-modality data (DISCOM). Our proposed DISCOM method includes two steps to find the optimal linear prediction of a continuous response variable using block-missing multi-modality predictors. In the first step, rather than deleting or imputing missing data, we make use of all available information to estimate the covariance matrix of the predictors and the cross-covariance vector between the predictors and the response variable. The proposed new estimate of the covariance matrix is a linear combination of the identity matrix, the estimates of the intra-modality covariance matrix and the cross-modality covariance matrix. Flexible estimates for both the sub-Gaussian and heavy-tailed cases are considered. In the second step, based on the estimated covariance matrix and the estimated cross-covariance vector, an extended
Lasso-type estimator is used to deliver a sparse estimate of the coefficients in the optimal linear prediction. The number of samples that are effectively used by DISCOM is the minimum number of samples with available observations from two modalities, which can be much larger than the number of samples with complete observations for all modalities. The effectiveness of the proposed method is demonstrated by theoretical studies, simulated examples, and a real application from the Alzheimer’s Disease Neuroimaging Initiative. The comparison between DISCOM and some existing methods also indicates the advantages of our proposed method.

Session 48: Statistical methods for biomarker data analysis

Systematic Evaluation of Statistical Methods in Subgroup Analysis
Xiaoqiang Wang and Yang Liu
University of Connecticut

Personalized medicine has become more and more popular in pharmaceutical industry with the development of subgroup analysis methodologies as its key. Searching for differences of treatment effects among subgroups is challenging because it is inherently a multiple testing problem with highly dependent test statistics as well as it is an underpowered study. While different paths have been explored for this topic, systematic comparison of the current existing methods for different scenarios in clinical trials is still an understudied topic. In this paper, we propose several evaluation criteria to investigate the advantages and disadvantage of different tree-based approaches for subgroup analyses. Extensive simulations studies are carried out to examine empirical performance of these methods.

Quantitative Reproducibility Analysis for Identifying Reproducible Targets from High-Throughput
Wenfei Zhang
Sanofi

High-throughput assays are widely used in biological research to select potential targets. Any single high-throughput experiment can efficiently study a large number of candidates simultaneously, but is subject to substantial variability. Therefore it is scientifically important to performance quantitative reproducibility analysis to identify reproducible targets, which are having consistent and significant signals across replicates studies. There are a few methods available but with limitations. We proposed a new reproducibility analysis method for identifying such reproducible targets. The proposed method is based on building Gaussian mixture model for the test statistics produced in replicate studies and selecting targets through their posterior probabilities. Both simulations studies and application examples are conducted to illustrate the performance of the proposed method in comparison to the existing methods. The proposed method is shown to be more accurate in selecting reproducible targets than the existing methods.

On Pharmacodynamic Biomarker for Early Decision Making
Atalanta Ghosh
Janssen Research and Development

Optimizing resources early in clinical development is more important than ever before. Maximizing the value of the pipeline is dependent upon efficient execution of programs along the phases of drug development, especially in early stages of the development during the learning phase. To make critical decision to fail fast is key to the success of optimal resource allocation. Biomarker plays a significant role in such early decision making process. The pharmacodynamic (PD) biomarkers are often used in early studies to understand the strength of the compounds potential success. Two such PD biomarkers, Mean Plasma Glucose (MPG) and Thrombin Time (TT) will be discussed in this presentation. Their definition and how they relate to the clinical investigation will be presented. Decision rules based on these biomarkers will be constructed using posterior probability. These decision rules and some properties of them will be discussed during the presentation.

Strategies for Clinical Development of Predictive Biomarkers
Glen Laird
Vertex Pharmaceuticals
glen_laird@VRTX.com

Frequently motivated by the drug mechanism of action, the incorporation of biomarkers into clinical drug development is becoming more common. However, there is little agreement on biomarker goals or strategy during different stages of development. Increasingly, there is interest in proactively incorporating at least a single biomarker into development strategy in a fashion that can impact pivotal studies as emerging data warrant. Recent developments in the area will be discussed including an overview of some publications.

Session 49: Big Data and Healthcare Analytics

Integrative Cox Regression for Modeling Uncertain Survival Records due to Imperfect Record Linkage
Wenjie Wang1, Robert Aseltine2, Kun Chen1 and Jun Yan1
1University of Connecticut
2University of Connecticut Health Center

In the era of big data, there has been an increasing need to use data integrated from disparate sources to conduct statistical analysis. The potential benefits from data integration, however, may be compromised by the induced data uncertainty due to incomplete/imperfect record linkage. Motivated by a suicide risk study, we consider survival modeling of uncertain event records arising from data integration. Specifically, a survival dataset derived from hospital discharge fails to capture all the events of death for patients, and the missing events may be recovered from a complete database of death records for a larger population. However, the original dataset can only be linked to the database by matching basic patient characteristics. As such, a censored patient from the original dataset could potentially be linked to a set of possible event times, which may or may not contain the truth. We develop an integrative Cox regression (ICox) approach, in which the uncertainty in event times is modeled probabilistically. The estimation procedure combines the ideas of profile likelihood and expectation conditional maximization (ECM) algorithm. Simulation studies demonstrate that under realistic settings of imperfect data linkage, the proposed method outperforms several competing approaches including multiple imputation. We apply the proposed approach in a marginal screening analysis for identifying risk factors associated with patient survival after suicide-related hospitalization in Connecticut. The identified diagnostics codes are consistent with existing literature and provide several new insights on suicide risk prediction and prevention.

Pragmatic Clinical Trials in Drug Development
Dingfeng Jiang
Pragmatic clinical trial (PCT) is an important type of study for effectiveness assessment in drug development. This session will introduce the general concept of PCT. The statistical aspects of PCT will be the focus of the discussion. An example of PCT will also be introduced for illustration purpose.

Big Data and Healthcare in Real World: Challenges and Opportunities
Yifan Xu
IBM Watson Health
ethan.yifanxu@gmail.com
The era of big healthcare data is here. With wider adoption of electronic healthcare records (EHR), healthcare systems continuously generate massive amount of EHR data every day. Big EHR data provide many benefits for statistical research such as (much) bigger sample size, larger number of observed metrics, more frequent measures in long time frames and relatively lower cost of obtaining the data. However, it also presents unique engineering, statistical, clinical and operational challenges.

In this talk, I will discuss these challenges and opportunities following the flow of EHR data in an analytical data pipeline. More specifically: Ingestion of EHR from multiple sources, data storage and transformation, handling of medical ontology, data quality, patient matching, selection bias and large scale risk modeling.

Angle based Multicategory Distance-weighted SVM
HUI SUN, Bruce Craig and Lingsong Zhang
Purdue University
sun400@purdue.edu
Classification is an important supervised learning technique with numerous applications. We develop an angle-based multicategory distance-weighted support vector machine (MDWSVM) classification method that is motivated from the binary distance-weighted support vector machine (DWSVM) classification method. The new method has the merits of both support vector machine (SVM) and distance-weighted discrimination (DWD) but also alleviates both the data piling issue of SVM and the imbalanced data issue of DWD. Theoretical and numerical studies demonstrate the advantages of MDWSVM method over existing angle-based methods.

Session 50: High-dimensional and Complex Data

Whiteout: Gaussian Adaptive Regularization Noise in Deep Neural Networks
Fang Liu1, Yinan Li2 and Ruoyi Xu2
1University of Notre Dame
2University of Science and Technology of China
fangliu.131@nd.edu
Noise injection (NI) is an off-the-shelf method to mitigate over-fitting in neural networks (NNs). The recent developments in Bernoulli NI as implemented in the dropout and shakeout procedures demonstrates the efficiency and feasibility of NI in regularizing deep NNs. We propose whiteout, a new regularization technique via injection of adaptive Gaussian noises into deep NNs. We show that whiteout is associated with a deterministic optimization objective function in generalized linear models with a closed-form penalty term which has connections with the bridge, lasso, ridge, and elastic net penalization; and it can be also extended to offer regularization similar to the adaptive lasso and group lasso regression. We also demonstrate that whiteout can be viewed as robust learning of NN model in the presence of small perturbations in input and hidden nodes. Compared to dropout, whiteout has better performance in training data of relatively small sizes with the sparsity introduced through the l1 regularization. Compared to shakeout, the penalized objective function in whiteout is more stable given the continuity of Gaussian noises. We establish theoretically that the noise-perturbed empirical loss function with whiteout converges almost surely to the ideal loss function, and the estimates of NN parameters obtained from minimizing the former loss function are consistent with those obtained from minimizing the ideal loss function. Computationally, whiteout can be incorporated in the back-propagation algorithm and is computationally efficient. The superiority of whiteout over dropout and shakeout in training NNs in classification is demonstrated using the MNIST and CIFAR-10 data.

High-dimensional robust regression
Po-Ling Loh
UW-Madison
loh@ece.wisc.edu
We present results for high-dimensional linear regression using robust M-estimators with a regularization term. We show that when the derivative of the loss function is bounded, our estimators are robust with respect to heavy-tailed noise distributions and outliers in the response variables, with the usual order of k log p/n rates for high-dimensional statistical estimation. Our results continue a line of recent work concerning local optima of nonconvex M-estimators with possibly nonconvex penalties, where we adapt the theory to settings where the loss function only satisfies a form of restricted strong convexity within a local neighborhood. We also discuss second-order results concerning the asymptotic normality of our estimators, and provide a two-step M-estimation algorithm for obtaining statistically efficient solutions within the local region.

Analyzing tree-based survival models: splitting rule, adaptive concentration and consistency
Yifan Cui1, Ruqing Zhu2, Mai Zhou3 and Michael Kosorok1
1University of North Carolina at Chapel Hill
2University of Illinois Urbana-Champaign
3University of Kentucky
rzhu@illinois.edu
As one of the most popular extensions of random forests, tree-based survival models lack established theoretical results and unified theoretical framework. We first investigate the method from the aspect of splitting rules, where the survival curves of the two potential child nodes are calculated and compared. We show that existing approaches lead to biased estimation of the within-node survival, and causes non-optimal selection of the splitting rules. This bias is due to the censoring distribution and the non-i.i.d. samples within each node. Based on this observation, we develop the adaptive concentration bound result for tree and forest versions of the survival tree models and established a general framework for showing the consistency. In particular, we show with two particular examples that, with some modification of the existing splitting rules, consistency results can be obtained. Interestingly, we also show that existing methods based on this biased selection of splitting rule can still lead to consistency as long as the censoring effect is weak.

Estimating Multi-level Brain Connectivity based on fMRI Data
Yumou Qiu1 and Xiao-Hua Zhou2
1University of Nebraska Lincoln
2Peking University
yumouqiu@unl.edu
As one of the most popular extensions of random forests, tree-based survival models lack established theoretical results and unified theoretical framework. We first investigate the method from the aspect of splitting rules, where the survival curves of the two potential child nodes are calculated and compared. We show that existing approaches lead to biased estimation of the within-node survival, and causes non-optimal selection of the splitting rules. This bias is due to the censoring distribution and the non-i.i.d. samples within each node. Based on this observation, we develop the adaptive concentration bound result for tree and forest versions of the survival tree models and established a general framework for showing the consistency. In particular, we show with two particular examples that, with some modification of the existing splitting rules, consistency results can be obtained. Interestingly, we also show that existing methods based on this biased selection of splitting rule can still lead to consistency as long as the censoring effect is weak.
Coherent activities between different brain regions play an important role in our cognition and behavior. The brain functional network is altered in neurodegenerative diseases, such as Alzheimer’s and Parkinson’s diseases. In this work, we proposed a hierarchical model on partial correlations to study the brain functional connectivity based on multi-subject functional magnetic resonance imaging (fMRI) data. A multiple testing procedure with false discovery proportion control is proposed to identify both the population and subject specific brain connectivity networks, which is adaptive to the temporal dependence among fMRI scans, and incorporates both the subject level estimation variation and among-subject heterogeneity variation. A computationally feasible algorithm is developed to implement the proposed procedure. Theoretical results and simulation studies demonstrate the good properties of the proposed procedure. A real example on fMRI data from normal healthy persons and patients with Parkinson’s disease shows that several within-group connections are missing in the brain network of Parkinson’s disease patients.

### Session 51: The state of some statistical challenges in Pharma 10 years hence

The state of some statistical challenges in Pharma 10 years hence

*Walt Offen, Frank Rockhold and Qi Jiang*

1 AbbVie Inc
2 Duke University
3 Amgen

Since the turn of the century, we have witnessed some exciting trends related to statistical aspects of clinical trial design, conduct, monitoring and analysis. Notions of adaptive clinical trial design, the use of Bayesian statistics for monitoring and designing trials, once the domain of small groups of innovators at the cutting edge, have become commonplace discussion topics. How we work has also changed dramatically. We are quick to turn to simulation to understand complexity. More have embraced working remotely for better or worse. Where are the trajectories of these established trends headed in 10 years? What are today’s nascent statistical trends and how might they change how we work? Our panel will provide some short opening remarks and then take on the role of soothsayers to field questions from the audience, opine about the future and offer recommendations for successfully navigating the changing landscape we find ourselves in as statisticians supporting drug development and health science.

### Session 52: Mei-Ling invited session 2

A New Measure of Synchronization to Quantify Brain Connectivity

Yang Zhou, Jane-Ling Wang, Hans-Georg Mueller and Owen Carmichael

1 University of California, Davis
2 Pennington Biomedical Research Center

When quantifying functional connectivity based on functional magnetic resonance imaging (fMRI) signals, the challenge is to assess the similarity of fMRI time courses that are observed for anatomically separated brain regions. The most prevalent measure is the temporal Pearson correlation (PC) but this static measure is less useful if connectivity fluctuates during a period of data collection. Dynamic functional connectivity has been proposed as an alternative to capture the dynamic features of functional connectivity. In this talk, we present a different approach that has a built-in dynamic feature but can be summarized by a static measure. The proposed measure is based on quantifying gradient synchronization by tracking concordance and discordance of the gradients between paired random curves. This gradient synchronization measure can be obtained by a simple procedure and consistency and asymptotic normality of the estimates towards a suitable target measure are derived under mild conditions. Our method is illustrated via simulations and with resting state blood oxygen level dependent (BOLD) fMRI signals from 20 specified hubs for both normal and Alzheimer’s patients.

### On Exact and Approximate Distributions of K-homopolymer for iid and Markov Dependent DNA Sequences

*James Fu*, Hsing-Ming Chang and Wendy Lou

1 University of Manitoba
2 National Cheng Kung University
3 University of Toronto

Walter.Offen@abbvie.com

Let $K = (k_1, k_2, k_3, k_4)$, $k_i \geq 0$, for $i = 1, 2, 3, 4$, be the lengths of simple homopolymers associated with $A_1 = A \ldots A$, $A_2 = B \ldots B$, $A_3 = C \ldots c$, $A_4 = D \ldots D$, respectively. For a given $K$, we refer the pattern $\Lambda_k = \{A_1, A_2, A_3, A_4\}$ as K-homopolymer. It is well-known that homopolymeric DNA subsequences are often associated with the accuracy of the alignment of DNA sequences. In this manuscript we study mainly the exact and approximate waiting distributions of K-homopolymer under the assumptions of independent identically distributed (iid) and first Markov dependent for the DNA sequences. Examples are given to illustrate our theoretical results.

### Statistical Analysis of Single Cell DNA methylation Data

*Yongsuk Park and Zhiguang Huo*

University of Pittsburgh

yongsuk@pitt.edu

DNA methylation plays an important role in normal cell and abnormal disease development such as cancer. It has been one of the most studied epigenetic markers. The detection technologies of DNA methylation have been shifted from mass spectrometry and array based assays to sequencing based technologies. Recent advances in single cell DNA methylation technologies enable us to understand dynamics of DNA methylation even further. However, up to now, there is no statistical method particularly considering the unique features of single cell DNA methylation. In this presentation, we present regional based beta-binomial model to analyze differential methylation and also propose entropy score to study the dynamics of DNA methylation.

### A Novel Statistical Model for Detecting DNA Methylation Marks Collected in Paired Design

Yunfeng Li, Jarrett Morrow, Benjamin Raby, Kelan Tantisira, Scott Weiss, Wei Huang and Weiliang Qiu

1 Zhejiang University
2 Harvard Medical School

A novel statistical model is proposed for detecting DNA methylation marks collected in paired design. Genetics plays important roles in complex human diseases. Many genetic effects are regulated by epigenetic mechanisms, such as DNA methylation, which will add a methyl group to DNA to regulate (usually inhibit) gene expression without changing genetic codes. By comparing the DNA methylation levels between dis-
eased subjects and healthy subjects, we can identify DNA methylation marks that are associated with complex human diseases. These differentially methylated marks will help us uncover the molecular mechanisms of complex human diseases. Cutting-edge technologies have been developed to measure the methylation levels for hundreds of thousands of DNA methylation marks simultaneously. However, technical noise and large between-sample variation could mask the true signal of DNA methylation. Paired design could efficiently reduce potential confounding effects. Probe-wise test (e.g. paired t-test for each DNA methylation mark) could be conveniently used to identify differentially methylated DNA methylation marks with good performance. A big hurdle for the probe-wise test is multiple testing that could result in huge false positive discovery. Therefore, p-values of the probe-wise tests have to be adjusted to control for false discovery rate. In this talk, we propose a novel model-based clustering model for detecting DNA methylation marks collected in paired design. The proposed clustering model does not involve hypothesis testing, hence, without multiple testing problems. Also, the proposed clustering model could explicitly utilize the mean vector structure of the within-pair difference of DNA methylation levels. Both simulation studies and real data analyses demonstrated good performance of the proposed model-based clustering model.

Session 53: Utilization of historical data into clinical trial development

The utility of collaboration: the TransCelerate Placebo/Standard of Care database
Jessica Lim
GSK

Clinical development modernization efforts have become essential as clinical trials have experienced increased expectations, costs, and design complexity. The utilization of historical clinical data can enhance drug research and development by refining study design, conduct, and analysis. TransCelerate is leading a collaboration across 14 companies to share placebo and standard-of-care (PSoC) clinical data with the aim to enhance innovative drug product development by better informing clinical safety interpretation and trial design. Use cases include developing a standing safety cohort for providing context around serious adverse events observed in ongoing trials, and using data from prior trials to reduce the number of patients in new proof of concept trials. This talk will cover use cases for the shared PSoC data, challenges around data sharing, successes to date, and important patient benefits.

A Bayesian Approach to Incorporate Historical Information in Clinical Trials

Judy Li, Wei-chen Chen and John Scott

A common question in clinical studies is how to use historical data from earlier studies, leveraging relevant information into the design and analysis of a new study. Bayesian approaches are particularly well-suited to this task, with their natural ability to borrow strength across data sources. In this talk, we introduce an eMAP approach for incorporating historical data into the analysis of clinical studies, and we discuss an application of this method to the analysis of observational safety studies for a class of products for patients with hemophilia A. The eMAP prior approach is flexible and robust to prior-data conflict. We conducted simulations to compare the frequentist operating characteristics of three approaches under different prior-data conflict assumptions and sample size scenarios.

Leveraging Historical Control Data: The Bayesian Augmented Control Method, with Case Studies and R Implementation
Hongtao Zhang, Saurabh Makhopadhyay, Qi Tang, Ran Liu, Kun Chen, Martin King and Bo Fu
1 AbbVie Inc.
2 Sanofi

Many clinical trials include a placebo or active control arm for which similar clinical data already exist from historical studies. In some situations, due to ethical, cost, or time considerations, it may be appealing to leverage the historical data for the control arm. The method, known as Bayesian Augmented Control (BAC), can help reduce the sample size and increase power. In this poster, a framework to determine the feasibility of applying the BAC method is presented. We showcase an R package that minimizes the programming effort to implement the BAC method. We illustrate the flow to apply BAC method in case studies.

Session 54: Recent advances in statistical genomics in personalized medicine

Integrating data-driven priors into GWAS and mediation analysis of GWAS with penalized regression
Sunduz Keles
University of Wisconsin, Madison

One of the contemporary challenges in understanding the results from genome-wide association studies is elucidating the potential roles of significant non-coding SNPs. Large consortia projects generated ample genomic and epigenomic data that are valuable for this task. We developed a number of statistical approaches for systematically incorporating such data-driven prior information into analysis of GWAS. Our approaches leverage penalized regression formulations of GWAS and mediation analysis. We provide large scale computational experiments that quantify when and how such information is useful as well as a theoretical exposition. Our analysis of several phenotypes from Framingham Heart Study illustrate the utility of this framework.

Statistical method for improving efficiency of CRISPR sgRNA design
Pei Fen Kuan, Scott Powers, Shuyao He, Kaiqiao Li, Xiaoyu Zhao and Bo Huang
1 Stony Brook University
2 Pfizer Inc.

CRISPR is a versatile gene editing tool which has revolutionized genetic research in the past few years. Optimizing sgRNA design to improve the efficiency of target/DNA cleavage is critical to ensure the success of CRISPR screens. By borrowing knowledge from oligonucleotide design and nucleosome occupancy models, we systematically evaluated candidate features computed from a number of nucleic acid, thermodynamic and secondary structure models on real CRISPR datasets. Our results showed that taking into account position-dependent dinucleotide features improved the design of effective sgRNAs with $AUC > 0.8$, and the inclusion of additional features offered marginal improvement. Using a machine-learning
Unbiased estimation of parent-of-origin effects using RNA-seq data from human
Vasyl Zhobotynsky\textsuperscript{1} and Wei Sun\textsuperscript{2}
\textsuperscript{1}UNC Chapel Hill
\textsuperscript{2}Fred Hutchinson Cancer Research Center
\texttt{vasyl@fredhutch.org}

RNA sequencing allows one to study allelic imbalance of gene expression, which may be due to genetic factors or genomic imprinting. It is desirable to be able to model both genetic and parent-of-origin effects simultaneously to avoid confounding and to improve the power to detect either effect. In a study of experimental cross, separation of genetic and parent-of-origin effects can be achieved by studying reciprocal cross of two inbred strains. In contrast, this task is much more challenging for outbred population such as human population. To address this challenge, we propose a new framework to combine experimental strategies and novel statistical methods. Specifically, we propose to collect genotype data from family trios as well as RNA-seq data from the children of family trios. We have developed a new statistical method to estimate both genetic and parent-of-origin effects from such data sets. We demonstrated this approach by studying 30 trios of HapMap samples. Our results support some of previous finding of imprinted genes and also recover some previously unknown imprinted genes.

Sparse additive index model for survival prediction with genomic data
Muxuan Liang, Kam-Wah Tsui and Sijian Wang
University of Wisconsin-Madison
\texttt{sijian.wang@wisc.edu}

In this talk, motivated by genomic studies with survival outcomes, we propose a sparse additive-index model to integrate pathway information to survival models. The method simultaneously constructs an index for each pathway and estimates the corresponding link function to connect the index to the outcome. A novel constraint is proposed to solve the identifiability issue when regularization on index parameters is present. Our proposed method can not only identify important pathways, but also select important genes within selected pathways. Furthermore, the proposed method has three good properties: 1) It is flexible to model the nonlinear association between genes and survival phenotype; 2) It automatically considers the interactions among genes within the same pathway; 3) It may distinguish the effects of a gene in all of pathways it belongs to. We have studied the theoretical properties of the methods. The methods are demonstrated using simulation studies and analysis on a TCGA ovarian cancer dataset.

Dynamic prediction involving high-dimensional factors based on the joint frailty-copula model
Takeshi Emura\textsuperscript{1}, Masahiro Nakatochi\textsuperscript{2}, Shigeaki Matsui\textsuperscript{3}, Hirofumi Michimae\textsuperscript{4} and Virginie Rondeau\textsuperscript{5}
\textsuperscript{1}National Central University
\textsuperscript{2}Nagoya University Hospital
\textsuperscript{3}Nagoya University Graduate School of Medicine
\textsuperscript{4}Kitasato University
\textsuperscript{5}INSERM, Universite de Bordeaux
takeshi_emura@gmail.com

The increasing availability of high-dimensional genetic factors and large-scale meta-analytic data sets for researchers has motivated the extension of traditional survival prediction based on the Cox proportional hazards model. The aim of our paper is to develop a risk prediction formula for patient survival according to genetic factors and dynamic tumour progression status based on meta-analytic data. To this end, we extend the existing joint frailty-copula model to a model allowing for high-dimensional genetic factors. Here, we suggest employing Tukey’s compound covariate to avoid the over-fitting of high-dimensional genetic factors, followed by Simon’s “P<0.001” criterion for the univariate tests. In addition, we propose a dynamic prediction formula to predict death given tumour progression events possibly occurring after treatment or surgery. We implement the computation software of the prediction formula in the joint.Cox R package. We also develop a tool to validate the performance of the prediction formula by assessing the prediction error. We illustrate the method with the meta-analysis of individual patient data on ovarian cancer patients.

Estimation of Strength of Detected Signals on Circadian Rhythm in Gene-Expression
Hua Yan Chen
University of Illinois at Chicago

In detecting genes controlling circadian rhythm, a $\chi^2$ test with two degrees of freedom is often employed. Since signals of circadian rhythm in gene-expression may not be strong enough to be detected at a high significance level required for adjusting for multiple tests, it is of substantial interest to estimate the total signal strength detected at a relatively low significance level. We extend the estimation approach for a univariate test to multivariate tests so that the total detected signal strength can be approximately unbiassedly estimated when $\chi^2$ tests with more than one degree of freedom are used for the signal detection. An approximate inference on the detected signal strength is also proposed. Simulation study demonstrates good performance of the proposed approach. The proposed approach is applied to the estimation of the total detected signals of circadian rhythm genes expressed in human brain tissue.

Tuning-free heterogeneity pursuit in massive networks
Zhao Ren\textsuperscript{1}, Yongjian Kang\textsuperscript{2}, Yingying Fan\textsuperscript{3} and Jinchi Lv\textsuperscript{2}
\textsuperscript{1}Zhejiang University
\textsuperscript{2}University of Wisconsin-Madison
\textsuperscript{3}National University of Singapore
zhao.ren@zju.edu.cn
Heterogeneity is often natural in many contemporary applications involving massive data. While posing new challenges to effective learning, it can play a crucial role in powering meaningful scientific discoveries through the understanding of important differences among sub-populations of interest. In this paper, we exploit multiple networks with Gaussian graphs to encode the connectivity patterns of a large number of features on the subpopulations. To uncover the heterogeneity of these structures across subpopulations, we suggest a new framework of tuning-free heterogeneity pursuit (THP) via large-scale inference, where the number of networks is allowed to diverge. In particular, two new tests, the chi-based test and the linear functional-based test, are introduced and their asymptotic null distributions are established. Under mild regularity conditions, we establish that both tests are optimal in achieving the testable region boundary and the sample size requirement for the latter test is minimal. Both theoretical guarantees and the tuning-free feature stem from efficient multiple-network estimation by our newly suggested approach of heterogeneous group square-root Lasso (HGSL) for high-dimensional multi-response regression with heterogeneous noises. To solve this convex program, we further introduce a tuning-free algorithm that is scalable and enjoys provable convergence to the global optimum. Both computational and theoretical advantages of our procedure are elucidated through simulation and real data examples.

Session 56: The Power of Integrative Genomics Data Analysis and Related Statistical Issues

A new algorithm for simultaneous clustering of genes in two species based on homology and expression

Yidan Sun and *Jingyi Jessica Li
UNIVERSITY OF CALIFORNIA, LOS ANGELES

In gene expression studies, clustering analysis is a key technique to identify functional gene modules. In the past few decades, many clustering algorithms have been developed to cluster genes in a species based on their expression values under different conditions. In comparative genomics, a common strategy is to link genes in different species by homology. An interesting question to ask is how to simultaneously identify gene clusters in two species by combining homology and gene expression information.

In this study, we propose a new algorithm based on bipartite spectral graph partitioning and tight clustering. We formulate a gene homology network between two species as a bipartite graph, with nodes on each side representing genes in each species and node covariates as gene expression values. Our goal is to identify tight and stable co-clusters of genes simultaneously in both species with strong homology and similar expression patterns. Compared with existing approaches with similar goals, our algorithm has two main advantages: (1) it does not rely on strong parametric assumptions, and (2) it does not force all nodes into clusters. We demonstrate the efficacy of our algorithm by comparing it with existing approaches on modENCODE gene expression data of two model organisms: D. melanogaster and C. elegans.

A statistical framework of mapping risk genes from de novo mutations in whole-genome sequencing study

*Xin He, Yiwen Liu, Ercument Cicek, Yanyu Liu, James

Analysis of de novo mutations (DNMs) from sequencing data of nuclear families has identified risk genes of a wide variety of complex diseases, including autism spectrum disorder (ASD). Such efforts have been mostly focused on mutations targeting protein-coding sequences. Non-coding regions harbor important regulatory information and contain the majority of the GWAS findings. Extending the DNM-based gene mapping approach to these non-coding regions is thus an important direction to advance our understanding of complex psychiatric disorders. Unlike coding sequences, however, we do not have a “genetic code” that allows us to easily predict the functional consequence of noncoding mutations, and as a result, how to interpret DNM data from whole genomes remains a major challenge. In this study, Dr. He and his colleagues developed a statistical framework for analyzing DNMs from whole-genome sequencing (WGS) data of parent-child trios. Built upon their earlier TADA model, the new model allows for combining both coding and non-coding mutations at the gene level to find causal genes. The model takes advantage of various genomic annotations, such as conservation scores and epigenomic marks, to prioritize non-coding mutations likely to be deleterious. An additional benefit of the model is that it easily accounts for batch effects (e.g. sequencing depth difference) and allows one to combine data from multiple studies. Application of their statistical framework to DNMs from WGS data of 300 ASD family trios (combining five studies) increases the power of disease gene mapping and predicts several new genes with possible roles in ASD.

De novo identification of functional elements in the DNA sequences via large-scale epigenetic data integration has broad utility in gene regulation and disease studies. Particularly, comparing functional elements across different cell types can reveal key regulators in a cell type specific context for understanding differential gene regulation and interpreting functions of DNA mutation that may impact phenotypes. The current epigenetic databases carry thousands of functional data sets in hundreds of human and mouse cell types, tissues and cell lines, yet the data sets for different epigenetic marks are often unmatched and unevenly distributed among cell types. A current solution for integrating these heterogeneous data sets are via epigenetic imputation, which is computationally expensive and subject to imputation errors. In this talk, I will present a new approach to leverage information across cell types for de novo identification of functional elements using all available epigenetic data sets without imputation. The approach has a much richer model structure than existing solutions and thus is more powerful. Simultaneously, the method is computationally fast even for analyzing hundreds of cell types jointly. The model can handle any incomplete combinations of epigenetic marks across cell types. As a special case, the method can use existing results in reference cell types to predict functional elements in new cell types with minimum data input.
Session 57: The Art of Evaluating the Surrogate Outcomes

Estimating Treatment Effects Using Multiple Surrogates: The Role of the Surrogate Score and Index
Hyunseung Kang
University of Wisconsin Madison
hyunseung@stat.wisc.edu

Estimating the long-term effects of treatments is of interest in many fields. A common challenge in estimating such treatment effects is that long-term outcomes are unobserved in the time frame needed to make policy decisions. One approach to overcome this missing data problem is to analyze treatments effects on an intermediate outcome, often called a statistical surrogate, if it satisfies the condition that treatment and outcome are independent conditional on the statistical surrogate. The validity of the surrogate condition is often controversial.

Here we exploit that fact that in modern datasets, researchers often observe a large number, possibly hundreds or thousands, of intermediate outcomes, thought to lie on or close to the causal chain between the treatment and the long-term outcome of interest. Even if none of the individual proxies satisfies the statistical surrogate criterion by itself, using multiple proxies can be useful and can yield efficiency gains. We focus primarily on a setting with two samples, an experimental sample containing data about the treatment indicator and the surrogates and an observational sample containing information about the surrogates and the primary outcome. We state assumptions under which the average treatment effect be identified on an intermediate outcome, which also has a positive ACE on the primary outcome, it is still possible that the treatment has a negative ACE on the primary outcome. Such a phenomenon is called the surrogate paradox and greatly challenges the use of surrogate. In this paper, we provide criteria to exclude the surrogate paradox for both the strong, and non-strong surrogates. Our criteria are optimal in the sense that they are sufficient and “almost necessary” to exclude the paradox: if the conditions are satisfied, the surrogate paradox is guaranteed to be absent while if the conditions fail, there exists a data generating process with surrogate paradox that can generate the same observed data. That is, our criteria capture all the information in the observed data to exclude the surrogate paradox rather than relying on unverifiable distributional assumptions.

Session 58: New Development on Matching Methodology for Causal Inference

Strong control of the familywise error rate in observational studies discovering effect modification
Jesse Hsu1, Jose Zubizarreta2, Dylan Small3 and Paul Rosenbaum1
1University of Pennsylvania
2Columbia University
hsu9@mail.med.upenn.edu

An effect modifier is a pretreatment covariate that affects the magnitude of the treatment effect or its stability. When there is effect modification, an overall test that ignores an effect modifier may be more sensitive to unmeasured bias than a test that combines results from subgroups defined by the effect modifier. If there is effect modification, one would like to identify specific subgroups for which there is the final outcome. This optimal surrogate has a number of highly desirable properties, including that the subgroup-specific average causal effect of treatment on the optimal surrogate is identical to the subgroup-specific average causal effect of treatment on the final outcome, which implies the optimal surrogate satisfies the Prentice (1989, Stat Med) definition of a valid surrogate endpoint for all subgroups defined by the baseline covariates collected in this trial. This in turn implies that the optimal surrogate is guaranteed to avoid the “surrogate paradox” where a negative treatment effect on the final outcome occurs despite promising positive results on the surrogate. Moreover, the optimality of the surrogate is robust to changes in the joint distribution of covariates, treatments, and candidate surrogate outcomes, providing a framework for estimation and inference of the treatment effect on the final outcome in a new trial that measures the optimal surrogate but not the final outcome. We present a targeted super-learner based estimator of this optimal surrogate and discuss the advantages of this estimator, including asymptotic efficiency for the current trial. The application of this methodology is illustrated using data from two randomized controlled dengue vaccine efficacy trials and through a simulation study.
evidence of effect that is insensitive to small or moderate biases. An exploratory method for discovering effect modification, and combine it with a confirmatory method of simultaneous inference that strongly controls the familywise error rate in a sensitivity analysis, despite the fact that the groups being compared are defined empirically, has been developed. In this talk, we will discuss a new form of matching, strength-k matching, permits a search through more than k covariates for effect modifiers, in such a way that no pairs are lost, provided that at most k covariates are selected to group the pairs. In a strength-k match, each set of k covariates is exactly balanced, although a set of more than k covariates may exhibit imbalance. We apply the proposed method to study the effects of the earthquake that struck Chile in 2010.

**Propensity Score Matching for Clustered Data**

*Mi-Ok Kim1, Bo Lu2 and Chunyan Liu3*

1 University of California, San Francisco
2 Ohio State University
3 Cincinnati Children’s Hospital Medical Center

Observational studies are a primary research method for comparative effectiveness research and patient-centered outcomes research. Their utility, however, is restricted by the fact that treatment choice is affected by known or unknown prognostic factors. This situation known as confounding by indication for treatment may render observational studies invalid and irrelevant unless properly addressed. Proper treatment of confounding is further complicated in data obtained from registries, network databases or the Electronic Health Record where subjects or patients are commonly clustered in ways that may be relevant to the analysis. We extend propensity score matching methodology to address measured and unmeasured confounding at the cluster level.

**Near-optimal matching algorithm for multi-group observational studies**

*Bo Lu*

The Ohio State University
[blu.232@osu.edu](mailto:blu.232@osu.edu)

Matching design is popular strategy to inferring causal relationship in observational studies, since it is less model dependent and provides interpretation similar to randomized studies. When there are two treatment arms, optimal algorithms exist for both bipartite and nonbipartite matching. Few studies, however, have used matching designs with more than two groups. This is due to the complexity of multiple-group matching algorithms and to the lack of optimal results. The implementation of the algorithm is not hard for a relatively small number of groups, i.e. 3-5. We propose an iterative multi-group matching algorithm that outperforms the nearest neighbor algorithm. We also discuss the extension of Rosenbaum’s sensitivity analysis to multi-group setup. We apply our proposed method to a real world health dataset for illustration.

**Optimal Multilevel Matching Using Network Flows: An Application to a Summer Reading Intervention**

*Samuel Pimentel1, Lindsay Page2, Matthew Lenard3 and Luke Keele3*

1 University of Pennsylvania
2 University of Pittsburgh
3 Wake County Public Schools
4 Georgetown University

[spimentel@gmail.com](mailto:spimentel@gmail.com)

Many observational studies of causal effects occur in settings with clustered treatment assignment. In studies of this type, treatment is applied to entire clusters of units. For example, an educational intervention might be administered to all the students in a school. We develop a matching algorithm for multilevel data based on a network flow algorithm. Earlier work on multilevel matching relied on integer programming, which allows for balance targeting on specific covariates but can be slow with larger data sets. Although we cannot directly specify minimal levels of balance for individual covariates, our algorithm is fast and scales easily to larger data sets. We apply this algorithm to assess a school-based intervention through which students in treated schools were exposed to a new reading program during summer school. In one variant of the algorithm, where we match both schools and students, we change the causal estimand to better maintain common support. In a second variant, we relax the common support assumption to preserve the causal estimand by only matching on schools. We find that the intervention does not appear to increase reading test scores. In a sensitivity analysis, however, we determine that an unobserved confounder could easily mask a larger treatment effect.

**Session 59: Recent Advances in Spatial Statistics: Theory and Application**

**Inference for spatial autocorrelation on a stream network**

*Dale Zimmerman*

University of Iowa
[dale-zimmerman@uiowa.edu](mailto:dale-zimmerman@uiowa.edu)

Recently, several models have been proposed for spatial dependence among observations taken at point locations on a stream network. Among these is the so-called "tail-up" model, under which only those observations from sites that are flow-connected (i.e., one is downstream of the other) are correlated; another is the "tail-down" model, which allows stream distance-dependent correlation between observations at any two sites in the network, even if they are flow-unconnected. I adapt the Diblasi-Bowman test for spatial independence in Euclidean geostatistics to test for tail-up dependence, using a scaled measure of "flatness" of the stream-distance semivariogram among flow-unconnected sites. A distinct test for tail-down dependence is also proposed. Distributional properties of the tests are investigated under the corresponding null hypotheses. The size and power of the tests are investigated by simulation. An example illustrates the use of the tests.

**Multivariate parametric models in ecology, with focus on addressing spatiotemporal variation**

*Brian Gray1 and Karl Oskar Ekvall2*

1 US Geological Survey
2 University of Minnesota

[brgray@usgs.gov](mailto:brgray@usgs.gov)

We review the use of multivariate parametric models in ecology, with focus on spatiotemporal variation. These models include multivariate generalized linear mixed models (MGLMMs), finite regression MGLMMs and latent variable models. These methods have seen little development for cluster sampling designs (e.g., where multivariate outcomes are obtained from each of multiple sites which, in turn, are nested within each of multiple lakes). Models of processes associated with clustered designs may be fitted using multilevel MGLMMs when multivariate outcomes are few. An alternative in this clustered or multilevel setting is to forego explicit inference on observation-level structure by using cluster-level sample means of the multivariate outcomes. Regardless of the selected approach, variance-covariance dimension reduction will of-
ten be important. We close by considering the use of a Kronecker decomposition for cases where spatiotemporal data have a two-dimensional structure. In a spatiotemporal setting, this decomposition assumes covariance functions from spatial (and temporal) blocks are proportional; we propose a relaxation of this decomposition to pertain to correlation rather than covariance functions. We motivate this work using dissolved oxygen sample means from sampling within 16 regions of the Upper Mississippi River over three seasons within each of 21 years.

Variogram models on spheres of all dimensions

Xintong Li\textsuperscript{1}, Juan Du\textsuperscript{2} and Chunseng Mu\textsuperscript{2}
\textsuperscript{1}Kansas State University
\textsuperscript{2}Wichita State University
\textsuperscript{xintongli@ksu.edu}

Variogram or variogram matrix functions play an important role in modeling dependence structure among multiple processes at different locations in spatial statistics. With more and more data collected on a global scale in environmental science, climatology, geophysics, and related fields, we focus on the characterizations of the variogram models on spheres of all dimensions for both stationary and intrinsic stationary, univariate and multivariate processes.

Some efficient approaches are proposed to construct a variety of variogram functions including simple polynomial structures, or to show whether an existing function on Euclidean space is valid on all spheres. In particular, the series representation and spherical behavior of intrinsic stationary random fields are elaborated in both theory and simulation study. We also use simulation and real data analysis to demonstrate the application of the proposed models and theoretical results involved in terms of estimation and kriging.

Flexible and efficient estimating equations for variogram estimation

Xiaohui Chang\textsuperscript{1}, Ying Sun\textsuperscript{2} and Yongtao Guan\textsuperscript{3}
\textsuperscript{1}Oregon State University
\textsuperscript{2}King Abdullah University of Science and Technology
\textsuperscript{3}University of Miami
\textsuperscript{xiaobull.chang@oregonstate.edu}

Variogram estimation plays a vastly important role in spatial modeling. Different methods for variogram estimation can be largely classified into least squares methods and likelihood based methods. In this paper, we propose a framework to estimate the variogram through a set of estimating equations. Our method is an alternative approach to likelihood based methods and includes commonly used least squares approaches as its special cases. This method is highly efficient as a low dimensional representation of the weight matrix is employed. We explore the statistical efficiency of various estimators constructed using this approach and examine the lag effect using numerical studies and simulations. The application of the proposed method to a hydrology dataset is also presented.

Session 60: Large-Scale Statistical Inference

On Polynomial Time Methods for Exact Low Rank Tensor Completion

Dong Xia and Ming Yuan
University of Wisconsin
\textsuperscript{myuan@stat.wisc.edu}

In this paper, we investigate the sample size requirement for exact recovery of a high order tensor of low rank from a subset of its entries. We show that a gradient descent algorithm with initial value obtained from a spectral method can, in particular, reconstruct a $d \times d \times d$ tensor of multilinear ranks $(r,r,r)$ with high probability from as few as $O(r^{5/2}d^{3/2} \log^{7/2} d + r^7d \log^7 d)$ entries.

Statistical and Computational Guarantees of Lloyd’s Algorithm and Its Variants

Ya Lu and Harrison Zhou
Yale University
\textsuperscript{hubin.zhou@yale.edu}

Clustering is a fundamental problem in statistics and machine learning. Lloyd’s algorithm, proposed in 1957, is still possibly the most widely used clustering algorithm in practice due to its simplicity and empirical performance. However, there has been little theoretical investigation on the statistical and computational guarantees of Lloyd’s algorithm. This paper is an attempt to bridge this gap between practice and theory. We investigate the performance of Lloyd’s algorithm on clustering sub-Gaussian mixtures. Under an appropriate initialization for labels or centers, we show that Lloyd’s algorithm converges to an exponentially small clustering error after an order of $\log n$ iterations, where $n$ is the sample size. The error rate is shown to be minimax optimal. For the two-mixture case, we only require the initializer to be slightly better than random guess.

In addition, we extend the Lloyd’s algorithm and its analysis to community detection and crowdsourcing, two problems that have received a lot of attention recently in statistics and machine learning. Two variants of Lloyd’s algorithm are proposed respectively for community detection and crowdsourcing. On the theoretical side, we provide statistical and computational guarantees of the two algorithms, and the results improve upon some previous signal-to-noise ratio conditions in literature for both problems. Experimental results on simulated and real data sets demonstrate competitive performance of our algorithms to the state-of-the-art methods.

A General Framework for Information Pooling in Two-Sample Sparse Inference

Yin Xia, Tony Cai and Wenguang Sun
\textsuperscript{1}Fudan University
\textsuperscript{2}University of Pennsylvania
\textsuperscript{3}University of Southern California
\textsuperscript{xiayin@fudan.edu.cn}

The conventional two-sample inference approach is based on the p-values of the differential data matrix and a selection of ranking cutoffs which adjust the multiplicity of comparisons. However, the data reduction step significantly loses the data information and the complicated dependence structures impose challenges on the data inference. In this talk, we introduce a general framework for information pooling of two-sample sparse inference that can handle a wide range of dependence structures. We construct primary statistics to directly assess the significance of the data difference and construct auxiliary covariates to assist inference indirectly by providing supplementary information. The proposed methodology involves three main steps: grouping, adjusting and pooling (GAP). Theoretical results guarantee the validity of the testing procedure. The new method can be widely applied to various settings of two-sample inference including simultaneous testing for high-dimensional sparse normal means, linear regressions, differential correlation analysis, and differential networks.

Session 61: Big Data and Interactions

TextM for Crowdsourcing in Epidemiology: Interface of Statistics and Computer Science

Jayang Sun, Katie Pezzort, Lu Wang and Rebecca Carter

\textsuperscript{1}University of Wisconsin
\textsuperscript{2}University of Pennsylvania
Crowdsourcing has become an important tool to rapidly amass huge amounts of data for scientific research. However, crowdsourcing approaches have not been standardized for application, say to epidemiology, and quality control has not been consistently used in crowdsourcing. In this work, we investigate challenges, trends, and classifications within Crowdsourcing in Epidemiology (CrowdEpi) and provide guidelines for an effective and responsive crowdsourcing protocol with quality control. To obtain data for CrowdEpi, we developed a Java-based text mining mining TextM and used an XML crawler to collect and query relevant articles from five main sources. To study the trends, we identified four main utilities within the CrowdEpi and developed a classifier (using SVM, LDA and PP) for automatically grouping articles into these four CrowdEpi areas and the nonCrowdEpi area. Our TextM can serve as a general-purpose text mining tool that allows uploading a specialized dictionary for a particular study. Our guidelines for crowdsourcing protocol should have applications beyond crowdsourcing in epidemiology. This work resulted from an interface of statistics and computer science for data science.

Hierarchical Sparse Modeling of Interactions
Xiaohan Yan1 and Jacob Bien2
1Cornell University, 2University of Southern California
xian@cornell.edu

Building predictive interaction models is an important yet challenging problem, especially in Big Data settings. To make this problem computationally tractable, it is common to demand that an interaction only be included in a model if both variables are marginally important. Many recently proposed methods enforce this requirement using group lasso regularizers with overlapping groups. We call attention to a little-known distinction between two types of overlapping group lasso regularizers, both of which have been used in the interaction modeling context. The interactions problem is a special case of what we call “hierarchical sparse modeling” (HSM). We provide a careful comparison – from both a computational and statistical perspective – of these two group lasso regularizers in the HSM context. This is joint work with Xiaohan Yan.

Boosting Gene Mapping Power and Efficiency with Efficient Exact Variance Component Tests of SNP Sets
Dandi Qiao1, Christoph Lange2, Nan Laird3, Sungho Won4, Craig Hersh5, Jarrett Morrow6 and Brian Hobbs1
1Brigham and Women’s Hospital
2Harvard T.H. Chan School of Public Health
3Seoul National University
reddaq@channing.harvard.edu

Whole-exome sequencing using family data has identified rare coding variants in Mendelian diseases or complex diseases with Mendelian subtypes, using filters based on variant novelty, functionality, and segregation with the phenotype within families. However, formal statistical approaches are limited. We propose a GEne-based SEGregation Test (GESE) that quantifies the uncertainty of the filtering approach. It is constructed using the probability of segregation events under the null hypothesis of Mendelian transmission. This test takes into account different degrees of relatedness in families, the number of functional rare variants in the gene, and their minor allele frequencies in the corresponding population. In addition, a weighted version of this test allows incorporating additional subject phenotypes to improve statistical power. We show via simulations that the GESE and weighted GESE tests maintain appropriate type I error rate, and have greater power than several commonly used region-based methods. We apply our method to whole-exome sequencing data of 114 subjects from 49 extended pedigrees with severe, early-onset chronic obstructive pulmonary disease (COPD) in the Boston Early-Onset COPD Study (BEOCOPD) and identify several promising candidate genes. Our proposed methods show great potential for identifying rare coding variants of large effect and high penetrance for family-based sequencing data. The proposed tests are implemented in an R package that is available on CRAN (https://cran.r-project.org/web/packages/GESE/).

Session 62: Recent Advances in Assessment of Agreement for Clinical and Lab Data

Bayesian Estimate of Concordance Correlation Coefficient
Dai Feng
Merck & Co., Inc
dai.feng@merck.com

Concordance correlation coefficient (CCC) is one of the most frequently used and reported scaled indices of agreement. There have been several frequentist methods proposed to estimate the CCC. Thus, several questions pertaining to the Bayesian estimate of the CCC naturally arose. In this talk, we aimed at answering the following questions: Is there any added value of a Bayesian approach when compared to frequentist methods? How is a Bayesian estimate of the CCC executed? Are there any potential limitations of...
Bayesian approaches and what are the possible solutions? Most of the Bayesian methods discussed in the entry were implemented in a package agRee available from the Comprehensive R Archive Network.

Methods for assessing the reliability of quality of life based on SF-36
Yi Pan
Centers for Disease Control and Prevention
[3NU58BCDC.GOV]
The 36-Item Short Form Health Survey (SF-36) has been widely used to measure quality of life. Reliability has been traditionally assessed by intraclass correlation coefficient (ICC), which is equivalent to Cronbach’s alpha theoretically. However, it is a scaled assessment of reliability and does not indicate the extent of differences because of measurement error. In our study, total deviation index (TDI) is used to interpret the magnitude of measurement error for SF-36, and a new formula for computing TDI for average item score is proposed. The interpretation based on TDI is simple and intuitive by providing, with a high probability, the expected difference that is because of measurement error. We also show that a high value of ICC does not always correspond to a smaller magnitude of measurement error, which indicates that ICC can sometimes provide a false sense of high reliability. The methodology is illustrated with reported SF-36 data from the literature and from real data in the Arthritis Self-Management Program.

Quantifying an Agreement Study
♦ Jason Liao and Jialin Xu
Merck
[jason.liao@merck.com]
In medical and other related sciences, clinical or experimental measurements usually serve as a basis for diagnostic, prognostic, therapeutic, and performance evaluations. Examples can be assessing the reliability of multiple raters (or measurement methods), assessing the suitability for tumor evaluation of using a local site or a central site in a randomized clinical trial (RCT), assessing the agreement of the clinical trial assay (CTA) and the in vitro diagnostic device (IVD) in the companion diagnostics (CDx) development for developing a personalized medicine, validating surrogate endpoints in a study, determining that the important outcome measurements are interchangeable among the evaluators in an RCT. Any elegant study design cannot overcome the damage by unreliable measurement. Many methods have been developed to assess the agreement of two measurement methods. However, there is little attention to quantify how good the agreement of two measurement methods is. In this talk, similar to the type I error and the power in describing a hypothesis testing, we propose quantifying an agreement assessment using two rates: the discordance rate and the tolerance probability. This approach is demonstrated through examples. An R-package was developed for this purpose and is used for demonstration.

Session 63: Integrating big and complex data with new statistical tools

SPATIAL-TEMPORAL GAUSSIAN STATE-SPACE MODELS
♦ Debashis Mondal and Chuxiao Wang
Oregon State University
[debashis@stat.oregonstate.edu]
In this talk, we present a scalable and matrix-free h-likelihood method for spatial-temporal Gaussian state-space models. The state vectors are assumed to follow spatial-temporal Gaussian autoregressions that are consistent with the conditional formulation of autonormal spatial fields. The h-likelihood method provides the same inference as that obtained from the Kalman filter and residual maximum likelihood analysis. However, for data from a large number of spatial sites, we show that our method has significant computational advantages. Furthermore, we provide details of the inference in small time steps and indicate how our method can be adapted to other complex spatial-temporal dynamical models based on stochastic partial differential equations. The method applies to data with both regularly and irregularly sampled spatial locations. We demonstrate the usefulness of our method with applications from environmental sciences and indicate some future directions.

Accounting for uncertainty in smoothing individual longitudinal profiles
♦ Joel Dubin and Shoja Chenouri
University of Waterloo
[jdubin@uwaterloo.ca]
In many applications for the analysis of individual longitudinal trajectories, a smoothing step takes place on the original, often sparsely collected, repeated measurements for each person, and the resulting individual curves are then used as the unit of analysis. Far too often, this smoothing step is ignored when ultimate inference is required in the utilization of these curves. We will describe different strategies for accounting for the uncertainty of the smoothing step in inference in the context of correlation of longitudinal curves, including a two-stage bootstrap approach and adjustments to this approach that allow for more robust inference. We will present both simulations to demonstrate the properties of these approaches and an application to a longitudinal dataset of serum protein measurements from hemodialysis patients.

Shape Constrained Inference in LASSO Regularized Regression
♦ Matus Maciak¹, Ivan Mizera² and Gabriela Ciuperca³
¹Charles University
²University of Alberta
³University Lyon
[maciak@karlin.mff.cuni.cz]
Complex data structures get a lot of attention in modern statistics and changepoints, or structural breaks respectively, definitely belong to this type of data structures. If, in addition, some qualitative restrictions are imposed on the final estimate, the whole estimation framework becomes a quite complex statistical problem. We present an innovative estimation approach based on the idea of various LASSO regularization techniques where the estimated changepoints are detected and estimated fully automatically, in just one single step. Moreover, the optional shape constraints can be usually expressed in terms of some simple linear restrictions which makes the estimation algorithm quite effective. Beside estimating the conditional mean of the target distribution only, we can also provide consistent estimates for the distributional quantiles which then offer more complex insight into the conditional distribution of interest. We provide some theoretical background on these types of models and the most important inference tools are discussed. The final sample performance is investigated using and extensive simulation study and some real data examples.
Session 64: Quantitative Collaborations and Partnerships with an Entrepreneurial Spirit

Quantitative Collaborations and Partnerships with an Entrepreneurial Spirit

Kelly Zou1, Alicia Toledo2, Aiyi Liu3, Birol Emir3 and Haoda Fu4
1Pfizer Inc
2Biostatistics Consulting, LLC
3National Institutes of Health
4Eli Lilly and Company, and Indiana University

This panel discussion session will highlight the challenges in and cutting-edge successful partnerships between academia and business, industry, and government organizations, with the co-sponsor of the symposium, the Committee on Statistical Partnerships among Academe, Industry and Government (SPAIG) of the American Statistical Association (ASA). The state-of-the-art statistical capability is typically a result of an entrepreneurial spirit. Several distinguished panelists will discuss (1) What do clients want? (2) Data quality and structures; (3) Communication skills. Sharing of ideas between quantitative analysts and scientists from different organizations leads to exchange visits, support for graduate students, consulting jobs, grant support for faculty, and continuing education opportunities for statisticians outside of academia. The intellectual exchange that results is a key component of such partnerships. We expect this session to have wide appeal, particularly given the increasing focus on interdisciplinary research and the emergence of opportunities for statisticians outside of academe. The intellectual exchange that results is a key component of such partnerships.

Session 65: The Jian-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences

Optimal combinations of biomarkers based on AUC and some recent developments for precision medicine

Xin Huang1, Gengsheng Qin2, Lu Tian3, Yixin Fang4 and Viswanath Devanarayan1
1AbbVie, Inc
2Georgia State University
3Stanford University
4New Jersey Institute of Technology

The area under the receiver operating characteristic curve (AUC) is a popular one-number accuracy summary index of a discriminatory system. When several biomarkers are available, one can combine them to achieve better diagnostic accuracy. In this presentation, we will introduce a nonparametric procedure to obtain the optimal linear combination that maximizes the AUC. Further, we will extend this procedure to the development of prognostic and predictive signatures for identifying patient subgroups of interest in the development of precision medicine.

Bayesian Inference for Network Meta-Regression Using Multivariate Random Effects

Hao Li1, Ming-hui Chen1 and Joseph Ibrahim2
1University of Connecticut
2University of North Carolina

Summary: Low-density lipoprotein cholesterol (LDL-C) has been identified as a causative factor for atherosclerosis and related coronary heart disease, and as the main target for cholesterol-lowering and lipid-lowering therapy. Statin drugs inhibit cholesterol synthesis in the liver and are typically the first line of therapy to lower elevated levels of LDL-C. On the other hand, a different drug, Ezetimibe, inhibits the absorption of cholesterol by the small intestine and provides a different mechanism of action. Many clinical trials have been carried out on safety and efficacy evaluation of cholesterol lowering drugs. To synthesize the results from different clinical trials, we examine treatment level (aggregate) network meta-data from 29 double-blind, randomized, active or placebo-controlled statins +/- Ezetimibe clinical trials on adult treatment-na ???.We patients with primary hypercholesterolemia. In this paper, we propose a new approach to carry out Bayesian inference for arm-based network meta-regression. Specifically, we group the variances of the random effects based on the clinical nature of treatments, and the determination of the number of groups and group membership is further guided by Bayesian model comparison criteria. The proposed approach is especially useful when some treatment arms are involved in only a single trial. In addition, a new Metropolis-within-Gibbs sampling algorithm is developed to carry out the posterior computations. In particular, the correlation matrix is generated from its full conditional distribution via partial correlations. The proposed methodology is further applied to analyze the network meta-data from 29 trials with 11 treatment arms.

Key words: Bayesian cumulative ranking curve; Deviance Information Criteria; Localized Metropolis algorithm; Logarithm of the Pseudo Marginal Likelihood; Partial Correlations.

Application of MC-SIMEX to log-logistic AFT models in survival analysis

Varadan Sevilimedu1, Lili Yu2, Hani Samawi1 and Haresh Rochani1
1Jiann Ping Hsu College of Public Health
2Jiann Ping Hsu College of Public Health

Survival analysis is the study of a living process before the occurrence of an event. Accelerated Failure Time models (AFT) serve as a useful tool in survival analysis to study the time of occurrence of an event and its relation to the covariates of interest. The accuracy of estimation of parameters in AFT models is dependent upon the correct classification of binary covariates. Considering that perfect classification is highly unlikely, it is imperative that the performance of the existing bias-correction methods be thoroughly analyzed in AFT models. However, certain areas of bias-correction in AFT models still remain unexplored. One of these unexplored areas is the situation where the survival times follow a log-logistic distribution. In this study, we evaluate the performance of the Misclassification simulation extrapolation (MC-SIMEX) procedure, a well-known procedure for bias-correction due to misclassification, in AFT models where the survival times follow a standard log-logistic distribution. In addition, a modified version of the MC-SIMEX procedure is also proposed, that provides an advantage in situations where the sensitivity and specificity of classification are unknown. Lastly, the performance of the original MC-SIMEX procedure in lung cancer data provided by the North Central Cancer Treatment Group (NCCTG), is also evaluated.

Privacy-Preserving Methods for Horizontally Partitioned In-
complete Data

Yi Deng, Xiaojian Jiang and Qi Long

Emory University
University of California
University of Pennsylvania

qz4z3@mail.missouri.edu

Distributed health data networks that leverage data from multiple sources have drawn substantial interests in recent years. However, such networks face challenges in the presence of missing data. The current state-of-the-art missing data methods require pooling data into a central repository before analysis, which may not be practical. In this paper, we propose a privacy-preserving framework for handling missing data. In this framework, each institution with a particular data source utilizes the local private data to calculate only necessary intermediate statistics which are then shared across all institutions for the proposed distributed algorithms. As such, the proposed framework as well as the involved distributed algorithms are privacy-preserving since no individual-level data are shared, leading to lower hurdles for collaboration across multiple institutions and stronger public trust with more institutions participating. To evaluate our proposed methods, we conduct simulation studies and then show the proposed privacy-preserving methods perform as well as the methods using the pooled data. We further apply the proposed methods on a dataset collected from 61 hospitals, for which we mimic the situation where the data are horizontally partitioned and each partition cannot be combined due to specific institutional policies.

Session 66: Recent advances in non-standard survival data

Current Status Data in the Presence of A Terminal Event
Lu Mao
University of Wisconsin-Madison
lmao@biostat.wisc.edu

We consider nonparametric estimation of current status data in the presence of a terminal event. This type of data often arises in biomedical studies, where the subject is scheduled to be tested for the presence of an asymptomatic non-terminal illness but may die before the test. We show that under this setting, neither the marginal distribution of the non-terminal event nor the sub-distributions are identifiable. Instead, we can only identify the distribution of the composite event, i.e., first occurrence of either the non-terminal or terminal events. We consider two estimators for the identifiable quantities, the maximum likelihood estimator (MLE) and a simpler maximum pseudo-likelihood estimator (MPLE). We cast the problem into the framework of order-restricted optimization, show existence and uniqueness of the estimators, and devise iterative convex minorant (ICM) algorithms for their computation. We perform simulation studies to assess and compare the finite sample performances of the proposed estimators.

Outcome-dependent sampling with interval-censored failure time data

Qingning Zhou, Jianwen Cai and Haibo Zhou

University of North Carolina at Chapel Hill
qz4z3@mail.missouri.edu

Epidemiologic studies and disease prevention trials often seek to relate an exposure variable to a failure time that suffers from interval-censoring. When the failure rate is low and the time intervals are wide, a large cohort is often required so as to yield reliable precision on the exposure-failure-time relationship. However, large cohort studies with simple random sampling could be prohibitive for investigators with a limited budget, especially when the exposure variables are expensive to obtain. Alternative cost-effective sampling designs and inference procedures are therefore desirable. We propose an outcome-dependent sampling (ODS) design with interval-censored failure time data, where we enrich the observed sample by selectively including certain more informative failure subjects like the case-cohort design. We develop a novel sieve semiparametric maximum empirical likelihood approach for fitting the proportional hazards model to data from the proposed interval-censoring ODS design. This approach employs the empirical likelihood and sieve methods to deal with the infinite-dimensional nuisance parameters, which greatly reduces the dimensionality of the estimation problem and eases the computation difficulty. The consistency and asymptotic normality of the resulting regression parameter estimator are established. The results from our extensive simulation study show that the proposed design and method works well for practical situations and is more efficient than the alternative designs and competing approaches. An example from the Atherosclerosis Risk in Communities (ARIC) study is provided for illustration.

Semiparametric Estimation of the Scale-Change Model with Panel Count Data under Informative Exams

Sy Han Chiou, Gongjun Xu, Jun Yan and Chiung-Yu Huang

Harvard University
University of Michigan
University of Connecticut
Johns Hopkins University
gongjun@umich.edu

Panel count data arise when recurrent events are the outcome of interest and when study subjects are observed periodically rather than continuously. As a result, instead of the exact event times, only the numbers of events that occur between successive examination times are observed. In many applications, the observation times can be informative about the underlying risk of recurrent events, violating the usual independent observation time assumption. In this talk, we consider semiparametric estimation of a semiparametric scale-change model for the underlying recurrent event process with panel count data, where the observation time process is allowed to be correlated with the underlying recurrent event process through a frailty variable. A novel estimation procedure for the regression parameters and the baseline rate function is proposed, which, in contrast to existing methods, is robust in the sense that it does not require the strong Poisson-type assumption for the underlying recurrent event process, nor does it requires a parametric assumption for the distribution of the unobserved frailty. Large-sample properties of the estimators are studied, and their variances are estimated by a model-based smoothed bootstrap procedure. Numerical studies demonstrated that the proposed point estimator and variance estimator perform well with practical sample sizes. The methods are applied to data from a skin cancer chemoprevention trial.
methods for left truncated survival data cannot be applied since no failure time is observed. We present some ongoing work on semi-parametric regression for current duration data.

Session 67: Recent advance in bioinformatics and computational biology

Personalized Risk Prediction for Glaucoma
X. Raymond Gao
University of Illinois at Chicago
rgao@uic.edu
Glaucoma is the leading cause of irreversible blindness worldwide, with primary open angle glaucoma (POAG) being the most common form of glaucoma. POAG usually does not produce typical symptoms for years and can be far advanced before patients notice an extensive visual field loss. Therefore, early detection is important. In recent years, polygenic risk scores (PRS) have been used in the risk prediction for many diseases, such as type 2 diabetes and coronary artery disease. Currently, few PRS studies have been reported in glaucoma. Furthermore, genetic studies in admixture populations, such as Hispanics/Latinos, remain a challenge. In this study, we construct PRS for POAG and its endophenotypes, build risk prediction models using advanced statistical approaches, such as LASSO regression, Xgboost and artificial neural network, and evaluate their risk-classification performance in Hispanic/Latino subjects. We carry out this precision-medicine study using genomics data from the Mexican American Glaucoma Genetic Study (MAGGS).

Integrating Diverse Genomic Data to Estimate Multiple Networks
Yuping Zhang
University of Connecticut
yuping.zhang@uconn.edu
Recent advances in high-throughput biotechnologies have generated various types of genetic, genomic, epigenetic, transcriptomic and proteome data across different biological conditions. It is likely that integrating data from diverse experiments may lead to a more unified and global view of biological systems and complex diseases. We present a coherent statistical framework for integrating various types of data from distinct but related biological conditions through graphical models. Specifically, our statistical framework is designed for modeling multiple networks with shared regulatory mechanisms from heterogeneous high-dimensional datasets. The performance of our approach is illustrated through simulations and its applications to cancer genomics.

A flexible R package to identify high quality gene modules from complex Omic networks
♦Xinyu Zhang1, Ying Hu2, Chunhua Yan2 and Ke Xu1
1Department of Psychiatry, Yale School of Medicine
2National Cancer Institute
xinyu.zhang
Network modularity analysis plays a critical role to understand complex disease-gene association and disease progression. In this challenge, we have developed a flexible R package for modularity analysis to map complex diseases. This R package has functions and visualization utilities to dissect complex networks, to identify modules, and to control quality (QC) of the identified modules. The module QC were performed using module weights and degree information. Multiple different sources of network raw data can be integrated together with normalization of gene-gene interaction weights.

An Isoform-free Model for Differential Expression Analysis in Large Sample RNA-seq Data
Yang Liu
Abbvie
yang.liu@abbvie.com
Next generation sequencing (NGS) technology has been widely used in biomedical research, particularly on those genomics-related studies. One of the NGS applications is the high-throughput mRNA sequencing (RNA-seq), which is usually applied to discover alternative splicing events, to evaluate gene expression level and to identify differentially expressed genes. Compared with the traditional microarrays, RNA-seq is more efficient and economical. Currently, many software tools have been developed for differential expression (DE) analysis, such as edgeR and DESeq; however, these methods are only sensitive to the total amount change of isoforms but insensitive to their proportion shift, which may reduce the accuracy of statistical inference. We developed and implemented a novel splicing-graph based negative binomial (SGNB) model for differential expression analysis in large sample RNA-seq data. The likelihood ratio test is used for finding DE genes. Computationally, we employed the expectation-maximization (EM) and the Newton-Raphson algorithms for parameter estimation. The main advantage of our method is that under some assumptions, it can detect not only the total amount change but the proportion shift without the pre-defined isoform structures, and therefore is expected to be more powerful and robust. We performed intensive simulations to compare our method with edgeR and DESeq. Under various scenarios we examined, the results showed that our model can achieve a better power, while correctly controlling the false discovery rate.

Session 68: Recent advances in statistical genetics

A powerful framework for integrating eQTL and GWAS summary data
Zhiyuan Xu1,* Chong Wu1, Peng Wei2 and Wei Pan1
1University of Minnesota
2MD Anderson Cancer Research Center
pan wei88@gmail.com
Two new gene-based association tests, called PrediXcan and TWAS for GWAS individual-level and summary data respectively, were recently proposed to integrate GWAS with eQTL data, alleviating two common problems in GWAS by boosting statistical power and facilitating biological interpretation of GWAS discoveries. Based on a novel reformulation of PrediXcan and TWAS, we propose more powerful gene-based association tests for single set or multiple sets of eQTL data. Our proposed methods are also applicable to GWAS summary statistics. The proposed methods are applied to several GWAS datasets, including two lipid summary association statistics based on 100,000 and 189,000 samples respectively, uncovering more and new trait-associated genes. The software implementing the proposed methods is freely available as an R package.

A novel test by testing an optimally weighted combination of variants with multiple traits
Zhenchuan Wang, * Qiyue Sha, Kai Zhang and Shuanglin Zhang
Michigan Technological University
qsha@mtu.edu
Pleiotropy, the effect of a single genetic variant or a single gene on multiple traits, has been a widespread phenomenon in complex diseases. Many large cohort studies collected a large number of correlated traits that can reflect underlying mechanism of pleiotropy. By joint analysis of multiple traits, we can not only gain more statistical
power to detect pleiotropic variants, but can also have a better understanding of the genetic architecture of the disease of interest. In this talk, I will present a test by testing an optimally weighted combination of variants with multiple traits (TOWmuT) to test association between multiple traits and a weighted combination of variants (both rare and common) in a genomic region. TOWmuT is applicable to different types of traits and can also incorporate covariates. I will show some of our simulation results for the performance of TOWmuT.

Dictionary learning based genotype imputation to improve power for association testing

• Xiaowei Wu and Hongxiao Zhu
Virginia Tech
xwuu0808@gmail.com

With the advances of high-throughput sequencing, imputing missing data in genotypes become increasingly critical for extracting and predicting informative signals of genetic association. Traditional genotype imputation methods rely on modeling allelic correlations among SNPs in a panel of known haplotypes, and then use this correlation information to extrapolate missing genotypes in the sample of interest. Due to heterogeneity among samples, such an approach may not achieve remarkably improved power for association testing. Here we consider unsupervised learning of the correlation structure in genotypes, and evaluating the power improvement achieved by this novel genotype imputation method. The proposed method is built on a probit model to link polychotomous genotypes with a set of latent variables. The mean of each latent variable is formulated as a sparse linear combination of a common over-complete basis, which constitutes a “dictionary”. We assume a beta-process prior for the coefficient sequence of the dictionary elements and develop a Markov chain Monte Carlo algorithm for posterior inference. Through simulations and real data analysis, we show that this dictionary learning based approach can effectively impute genotypes without learning allelic correlations from reference panel and/or using individual relatedness from known pedigree, and thus provides an potential solution to boost power in genetic association studies.

Session 69: Urging a paradigm change: BFF inferences in the era of data science

On the cross correlations under high dimension

Han Xiao
Rutgers University
hxiao@stat.rutgers.edu

As an initial step before modeling high dimensional time series, it is of interest to check whether the component series are correlated. We suggest to perform the test based on the sample cross correlations of the original series, under the presence of temporal dependence. We consider test statistics based on: maximum sample cross correlations, maximum of the pairwise portmanteau type statistics, and some other variants. Asymptotics are developed in the high dimensional setting where the dimension p can grow either as a power of the sample size T, or as an exponential function of T. We employ the moving blocks bootstrap method to calibrate the sizes of the tests for finite samples. Extensions to nonstationary time series are also considered.

Non-penalized variable selection in high-dimensional linear model via generalized fiducial inference

Jonathan Williams and Jan Hannig

University of North Carolina at Chapel Hill
jan.hannig@unc.edu

Standard penalized methods of variable selection and parameter estimation rely on the magnitude of coefficient estimates to decide which variables to include in the final model. However, coefficient estimates are unreliable when the design matrix is collinear. To overcome this challenge an entirely new method of variable selection is presented within a generalized fiducial inference framework. This new procedure is able to effectively account for linear dependencies among subsets of covariates in a high-dimensional setting where p can grow almost exponentially in n, as well as in the classical setting where p > n.

It is shown that the procedure very naturally assigns small probabilities to subsets of covariates which include redundancies by way of explicit L0 minimization. Furthermore, with a typical sparsity assumption, it is shown that the proposed method is consistent in the sense that the probability of the true sparse subset of covariates converges in probability to 1 as n and p go to infinity. Very reasonable conditions are needed, and little restriction is placed on the class of 2p possible subsets of covariates to achieve this consistency result.

Higher-Order BFF

Todd Kuffner
Washington University in St. Louis
kuffner@wustl.edu

Large-sample agreements to first order O(n−1/2) in the available sample size n, of Bayesian, fiducial, marginal frequentist, and conditional frequentist inferences for low-dimensional parameters of interest, is a beautiful and encouraging phenomenon. It is inherently desirable in applied statistics that the inferences for the same parameters, when derived from competing inferential paradigms, should yield similar conclusions. One does not want the conclusions of a statistical analysis to depend heavily on the philosophical convictions of the data analyst. In some sense it is not surprising that first-order agreement is generally possible in parametric settings. But higher-order agreement, in the sense of agreement to O(n−1) or better, is elusive and difficult to characterize. Before discussing some available results in this direction, I consider the natural question: does it matter? That is, why is higher-order agreement important? Examples are used to illustrate the importance of higher-order results in the practically relevant setting of small- to moderate-sample sizes, or even in large samples with very high-dimensional parameters. Then some explanations are offered for the possibility of higher-order reconciliation of these competing inferential paradigms.

An Objective Prior for Hyperparameters in Normal Hierarchical Models

James Berger1, Dongchu Sun2 and Chengyuan Song3
1Duke University
2University of Missouri
3East China Normal University
sund@missouri.edu

Hierarchical models are the workhorse of much of Bayesian analysis, yet there is uncertainty as to which objective priors to use for hyperparameters. Formal approaches to objective Bayesian analysis, such as the Jeffreys’s-rule approach or reference prior approach, are only implementable in simple hierarchical settings. It is common to use less formal approaches, such as utilizing formal priors from non-hierarchical models in hierarchical settings. This can be fraught with danger, however. For instance, non-hierarchical Jeffreys’s-rule priors for variances or covariance matrices result in improper posterior distributions if they are used at higher levels of a hierarchical
model. Berger et al. (2005) approached the question of choice of hyperpriors in normal hierarchical models by looking at the frequentist notion of admissibility of resulting estimators. The motivation was that hyperpriors that are too diffuse result in inadmissible estimators, while hyperpriors that are concentrated enough result in admissible estimators. Hyperpriors that are ‘on the boundary of admissibility’ are sensible choices for objective priors, being as diffuse as possible without resulting in inadmissible procedures. The admissibility (and propriety) properties of a number of priors were considered in Berger et al. (2005), but no overall conclusion was reached as to a specific prior to recommend, in part because they were not able to prove admissibility for the leading candidate prior. In this talk, we complete the story and propose a particular objective prior for use in normal hierarchical models, based on considerations of admissibility, ease of implementation (including computational considerations), and performance.

Session 70: Panel sessoin on leadership

First and Foremost... Know Thyself
Chito Hernandez
Biomarin
cbernehandez@bmrn.com

The title is as intriguing as the question of what makes you a good statistician. We can all dig into very specific actions so that you can break free from how you have been thinking about yourself and your work and what it would take for you to get to the next level. That next level will vary depending on where you are in your career. But we can all figure it out together.

Session 71: General Developments in Nonclinical Statistics

Bayesian hierarchical model estimation and comparison of immunogenicity assay cut-points from pre-st
• Dave LeBlond1, Rong Zeng2, Lu Xu2 and Robert Singer3
1Statistical Consultant
2OncoMed Pharmaceuticals
3Robert Singer Consulting
david.leblond@sbcglobal.net

Therapeutic proteins may elicit immune reactions resulting in the appearance of anti-drug antibodies (ADA) in sera from treated subjects. ADA may pose efficacy or safety risks so their presence in clinical trial subjects is monitored using screening and confirmatory assays. The presence of ADA is suspected when an assay signal exceeds a specific cut-point estimated from pre-study normal human sera or in study baseline responses. Assay cut-point determination is an area of active research but must conform to regulatory guidance. Current practice employs analytical/biological outlier determination and distributional assumptions to assure a minimum specific false positive rate with high confidence. Comparisons among pre- and in-study subject sera populations are periodically made to assess the need for study specific cut points (SSCP). Recently, Yang and colleagues developed Bayesian hierarchical models to separate biological and analytical variation and non-normal distributional models (e.g., skew-t) of the biological component. Here we extend these developments to include the use of Rosner’s multiple-outlier test, a Johnson SU distributional model of biological variability. We show how to estimate the cut point as a conservative lower bound on a quantile of the posterior predictive (PP) distribution of future signals. We also show how to compare PP distributions from different studies as an aid to risk-based SSCP decision making. We will illustrate our approach with actual study data and calibrate its frequentist properties using Monte-Carlo simulation.

Use of High Confidence Sampling Plans for Product Assessment
Mark Johnson
Abbvie Inc.
mark.d.johnson@abbvie.com

This presentation will discuss use of attribute (e.g., pass or fail assessment) and variables (e.g., quantitative statistical assessment) sampling and inspection approaches for use during product process validation activities, as well as use to address impact of process excursions or limit violations on quality of manufactured product.

Exact Trend Test on Comparing Tumor Incidence in Transgenic Mouse Carcinogenicity Studies
Lei Shu
Astellas Pharma
lei.shu@astellas.com

Statistical positive trend tests to predict drugs’ tumorigenic potential for long-term (usually 2 years) carcinogenicity studies, for example, Peto’s test, are well accepted in practice and by regulatory agencies. Recently short-term (usually 6-month) transgenic mouse studies become widely used as an alternative to lifetime mouse carcinogenicity studies. In general the tumor incidence rates in a short-term study are much lower than those for a traditional long-term study, so exact positive trend test should be used. However, little is discussed about exact statistical trend tests for transgenic mouse studies. In this presentation, we evaluated three types of exact methods for testing tumor trend via simulation: exact Cochran-Armitage (C-A) test, exact Peto’s test and a proposed conditional exact Poly-k test. Based on an original exact Poly-k test introduced by Rahman that requires extensive computation, we proposed an exact Poly-k test that modifies the original exact test by conditioning on observed mortality data. The proposed conditional exact Poly-k test significantly reduces the computational complexity and can be implemented in practice.

Disclaimer: The speaker is a paid employee of Astellas. This presentation is intended for informational purposes only and does not replace independent professional judgment. This presentation is not intended to be legal advice. Statements of fact, positions taken and opinions expressed are those of the speaker individually and, unless expressly stated to the contrary, do not necessarily reflect the opinion or position of the speaker’s employer, Astellas, or any of its subsidiaries and/or related entities.

A machine learning approach to biomarker discovery from high dimensional omics data
Jie Cheng
Abbvie Inc.
jie.cheng@abbvie.com

We will present a machine learning based feature selection method for high dimensional omics data. By performing grid search using a pair of statistics, this method can automatically adapt to the characteristics of different datasets and parsimoniously select the most informative biomarkers. Several real world datasets will be used to demonstrate the effectiveness of this method. Furthermore, we will suggest good practice guidelines for feature selection from omics data.
Session 72: Topics on statistical modeling and inference (I)

A unified mixture regression approach to covariate misclassification and missingness
♦ Michelle Xia1 and P. Richard Hahn2
1 Data Science and Statistics, AbbVie Inc.
2 Booth School of Business, University of Chicago

cxia@niu.edu

This paper considers the problem of mismeasured and missing categorical covariates in the context of regression modeling; if unaccounted for, such misclassification is known to result in misestimation of model parameters. Here, we exploit the fact that explicitly modeling covariate misclassification and missingness leads to a mixture representation. Assuming common parametric families for the mixture components, and assuming that the misclassification and missingness occurrence is independent of the response variable, the mixture representation permits model parameters to be identified even when misclassification probabilities are unknown. Previous approaches to covariate misclassification use multiple surrogate covariates and/or validation data on the magnitude of errors. Based on this mixture structure, we demonstrate that valid inference can be performed on all the parameters even when no such additional information is available. Using Bayesian inference, the method allows for learning from data combined with external information on the magnitude of errors when such information does become available. The method is applied to adjust for misclassification on self-reported cocaine use in the Longitudinal Studies of HIV-Associated Lung Infections and Complications (Lung HIV). We find a substantial and statistically significant effect of cocaine use on pulmonary complications measured by the relative area of emphysema, whereas a regression that does not adjust for misclassification yields a much smaller estimate.

Robust Fitting of Mixture Regression Models
Xiuqin Bai
xbai@ewu.edu

We propose a new robust mixture of regression models. The existing methods for fitting mixture regression models assume a normal distribution for error and then estimate the regression parameters by the maximum likelihood estimate (MLE). In this project, we demonstrate that the MLE, like the least squares estimate, is sensitive to outliers and heavy-tailed error distributions. We propose a robust estimation procedure and an EM-type algorithm to estimate the mixture regression models. Using a Monte Carlo simulation study, we demonstrate that the proposed estimation method is robust and works much better than the MLE when there are outliers or the error distribution has heavy tails. In addition, the proposed robust method works comparably to the MLE when there are no outliers and the error is normal.

Impact of variance-covariance specification in MMRM on multiple comparisons inference
♦ Meng Li and Jason Hsu
Ohio State University
mengli2011@gmail.com

Clinical trials in diabetes, schizophrenia, and Alzheimer’s Disease measure effects of treatments on each patient repeatedly at multiple time points. Such data are typically analyzed using a mixed effects model. Insight into the variance-covariance of the Repeated Measures is given by deriving and examining its structure for a simple Random coefficients response model. To isolate the consequence of specifying a variance-covariance structure far from the truth, we cast our investigation in a setting where the fixed treatment effect is orthogonal to the random effect and the interactions. By analyzing data sets satisfying such orthogonality condition under different variance-covariance types, we find the type specification itself has a significant impact on testing for treatment efficacy at multiple time points.

A Gatekeeping Testing Procedure for the Statistical Assessment of a Second Generation AD Product
Ling Chen
ling.chen@fda.hhs.gov

In April 2015 FDA issued the final guidance to assist pharmaceutical companies in developing opioid products with abuse-deterrent (AD) properties. Since 2010, several opioids with AD formulation technology have been approved by the FDA, and more are in development. The question has been raised regarding how to compare a second generation AD product (a test product) to an approved AD version of the same opioid product (an approved AD product) in a clinical abuse potential study. There have been proposals not to include an immediate release (IR) or Non-AD extended release (ER) opioid product as a positive control in the study, and to compare a test product to an approved AD product using a non-inferiority test. This presentation discusses the reasons why an IR or Non-AD ER opioid product should be included in such a study as a positive control, and when a second generation AD product is proposed, why one should not compare it to an approved AD product using a non-inferiority test, and then proposes a gatekeeping testing procedure for the comparison.

Variable selection for random effects two-part model
♦ Dongxiao Han1, lei Liu2, xiaogang Su3 and Bankole Johnson4
1 Northwestern University
2 Northwestern University
3 University of Texas at El Paso
4 University of Maryland
dongxiao han@northwestern.edu

Zero-inflated continuous (or semi-continuous) data are commonly encountered in longitudinal biomedical studies. The presence of a large portion of zero values and the right skewness for the positive continuous values are distinct features of such data, which are often tackled by the two-part model. More recently, random effects two-part models have been used to study longitudinal zero-inflated continuous data (Olsen and Schafer 2001; Tooze, Grunwald and Jones, 2002). An additional complicating issue in personalized medicine is the presence of large number of genetic and other covariates. In this paper we consider variable selection in repeated measures of zero-inflated data. We apply the minimum information criterion method (Su et al., 2016) to select variables in the random effects two-part model. The estimation is conducted by adaptive Gaussian quadrature which can be conveniently implemented in SAS Proc NLMIXED. The behavior of our approach is evaluated through simulation studies, and an application to a longitudinal alcohol dependence study is provided.

Keywords: Variable selection; High dimensional; Mixed effects; Precision medicine; Tuning parameter; Pharmacogenetics.

Session 74: Statistical and Computational Challenges for Single-cell Sequencing

SCnorm: A quantile-regression based approach for robust normalization of single-cell RNA-seq data
♦ Rhonda Bacher1, Li-Fang Cha2, Audrey Gasch1, James Thomson2, Ron Stewart1, Michael Newton1 and Christina
Normalization of RNA-sequencing data is essential for accurate downstream inference, but the assumptions upon which most methods are based do not hold in the single-cell setting. Consequently, applying existing normalization methods to single-cell RNA-seq data introduces artifacts that bias downstream analyses. To address this, we developed SCNorm to enable efficient and accurate scRNA-seq normalization. Simulation and case study results suggest that the framework provides for increased accuracy in fold-change estimation as well as improvements in downstream inference.

Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning

Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti and Seraphim Batzoglou
Stanford University
bowang87@stanford.edu

Single-cell RNA-seq technologies enable high throughput gene expression measurement of individual cells, and allow the discovery of heterogeneity within cell populations. Measurement of cell-to-cell gene expression similarity is critical to identification, visualization and analysis of cell populations. However, single-cell data introduce challenges to conventional measures of gene expression similarity because of the high level of noise, outliers and dropouts. Here, we propose a novel similarity learning framework, SIMLR (single-cell interpretation via multi-kernel learning), which learns an appropriate distance metric from the data for dimension reduction, clustering and visualization applications. Benchmarking against state-of-the-art methods for these applications, we used SIMLR to re-analyse seven representative single-cell data sets, including high-throughput droplet-based data sets with tens of thousands of cells. We show that SIMLR greatly improves clustering sensitivity and accuracy, as well as the visualization and interpretability of the data.

Global Prediction of Chromatin Accessibility Using RNA-seq from Single Cell and Small Number of Cell

WeiQiang Zhou, Zhicheng Ji and Hongkai Ji
Department of Biostatistics, Johns Hopkins Univ.
hji@jhu.edu

Conventional high-throughput technologies for mapping regulatory element activities such as ChIP-seq, DNase-seq and FAIRE-seq cannot analyze samples with small number of cells. The recently developed ATAC-seq allows regulome mapping in small-cell-number samples, but its signal in single cell or samples with ??500 cells remains discrete or noisy. Compared to these technologies, measuring transcriptome by RNA-seq in single-cell and small-cell-number samples is more mature. Here we develop BIRD, Big Data Regression for predicting DNase I hypersensitivity using gene expression. We show that BIRD can globally predict chromatin accessibility and infer regulome using RNA-seq. Genome-wide chromatin accessibility predicted by RNA-seq from 30 cells is comparable with ATAC-seq from 500 cells. Predictions based on single-cell RNA-seq can more accurately reconstruct bulk chromatin accessibility than using single-cell ATAC-seq by pooling the same number of cells. Integrating ATAC-seq with predictions from RNA-seq increases power of both methods. Thus, transcriptome-based prediction can provide a new tool for decoding gene regulatory programs in small-cell-number samples.
for analysis. In this paper, we apply a kernel RV coefficient (KRV) test to evaluate the overall association between host gene expression and microbiome composition. The KRV statistic can capture nonlinear correlations and complex relationships among the individual data types and between gene expression and microbiome composition through measuring general dependency. Testing proceeds via a similar route as existing tests of the generalized RV coefficients and allows for rapid p-value calculation. Strategies to allow adjustment for confounding effects, which is crucial for avoiding misleading results, and to alleviate the problem of selecting the most favorable kernel are considered. Simulation studies show that KRV is useful in testing statistical independence with finite samples given the kernels are appropriately chosen, and can powerfully identify existing associations between microbiome composition and host genomic data while protecting type I error. We apply the KRV to a microbiome study examining the relationship between host transcriptome and microbiome composition within the context of inflammatory bowel disease and are able to derive new biological insights and provide formal inference on prior qualitative observations.

Trace Evidence Analysis Utilizing Metagenome Sequence Data

Kyle Carter, Meng Lu and Lingling An
University of Arizona

Metagenomic analysis is a promising alternative to traditional forensic methods when analyzing trace evidence samples. However, the microbial communities found on evidence samples usually contain microbes from multiple human and environmental contributors. When there are multiple contaminants, traditional source tracking methods struggle to distinguish true microbial contaminant sources from similar populations. We propose a filtered microbial source tracking method that utilizes ecological dissimilarity to select true microbial contaminant sources and estimate the proportion of microbes in the evidence sample that derives from the selected contaminants.

A Marginalized Two-Part Beta Regression Model for Microbiome Compositional Data

Haitao Chai, Hongmei Jiang, Lu Lin and Lei Liu
Shandong University, Northwestern University

Human microbial communities are associated with many human diseases. One goal of human microbiome studies is to detect differential abundance across clinical conditions and treatment options. However, the microbiome compositional data (denoted by relative abundance) are highly skewed, bounded in [0, 1), and often with many zeros. A two-part model is commonly used to separate zeros and positive values explicitly by two submodels: a logistic model for the probability of a specie being present in Part I, and a Beta regression model for the relative abundance conditional on the presence of the specie in Part II. However, the regression coefficients in Part II cannot provide an unconditional (marginal) interpretation of covariate effects on the microbial abundance, which is of great interest in many applications. In this paper, we propose a marginalized two-part Beta regression model which captures the zero-inflation and skewness of microbiome data and also allows investigators to examine covariate effects on the unconditional (marginal) mean. We demonstrate its practical performance using simulation studies and apply the model to an Inflammatory Bowel Disease (IBD) study. We find significant treatment effects on the marginal mean of relative abundance of three genera, while the treatment effects are not significant in either Part I or Part II of the traditional two-part models.

Session 76: Recent Advancements in Personalized Medicine

Empirical Likelihood for Comparing Two Survival Functions

Hsin-wen Chang
Academia Sinica

Comparing two survival functions is a fundamental problem in many clinical applications. Despite the classical procedures like the log-rank test, there has been increasing evidence that empirical likelihood can provide more powerful hypothesis tests. In this talk we introduce R functions for empirical likelihood based hypothesis tests. Both two-sided and one-sided tests are provided. The software will be made freely available in the R package EL2Surv. Simulation results demonstrating the superiority of these procedures over traditional survival methods are presented. The use of package EL2Surv is also illustrated through different applications.

Subgroup identification with latent Dirichlet allocation

Hyonho Chun
Purdue University

It is a general belief that one global model may not be suitable to describe diverse drug responses. There may be a specific subgroup (individuals) that reacts especially well to a certain type of drug than the others. It is then very important to identify these subgroups that share genomic features as well as drug responses. We will apply the latent Dirichlet allocation model to the cancer cell line encyclopedia datasets to identify these subgroups.

Adjusting treatment effect estimated from an exploratory biomarker analysis using resampling methods

Aimee Wang, Tianle Hu and Honglu Liu
Eli Lilly and Company

There has been growing interest in subgroup identification based on a continuous biomarker in oncology clinical trials. Considerable amount of research has been devoted to determination of the optimal cut-off point of the biomarker accounting for multiplicity. However, little is focused on adjusting the raw treatment effect from the initial subgroup for properly powering the subsequent confirmatory trial. In this presentation, we discuss the use of resampling methods to adjust the treatment effect estimates based on post-hoc exploratory subgroup analyses, when the trial overall is negative and the best subgroup is selected to move forward in a post-hoc manner. We show that our proposed estimator properly accounts for selection bias. We then examine the operating characteristics of the confirmatory trials designed based on the adjusted effect size and compare them with those of trials designed based on unadjusted effect size, to demonstrate the importance of bias adjustment given the post-hoc nature of the initial subgroup identification.

Session 77: Statistical Tests for Data Measured on Taxonomic Trees in Microbiome Studies

A General Framework for Association Analysis of Microbial Communities on a Taxonomic Tree

Zheng-Zheng Tang, Guanhua Chen, Alexander Alekseyenko and Hongze Li

2017 ICSA Applied Statistics Symposium, Chicago, June 25-28
The phylogenetic LASSO and the microbiome

Stephen T Rush\textsuperscript{1}, Jessmyn Niergarth\textsuperscript{2} and Peter T Kim\textsuperscript{1}

\textsuperscript{1}University of Guelph
\textsuperscript{2}Fred Hutchinson Cancer Research Center

Scientific investigations that incorporate next generation sequencing involve analyses of high-dimensional data where the need to organize, collate and interpret the outcomes are pressing important. Data of the microbiome can currently be collected, leading to possible advances in personalized medicine. In this thesis, we lay down a statistical framework for incorporating metagenomic information in predictive modeling with a view toward synthesis of products tailored to individual patients. In particular, we develop the phylogenetic LASSO (-LASSO), a form of model regularization which incorporates known relationships between predictors for the purpose of model selection. We apply the -LASSO to a pilot metagenomic study on the efficacy of fecal microbiota transplantation in the treatment of Clostridium difficile infections. Although the thesis applies the technique to data for a particular infectious disease, the methodology is sufficiently rich to be expanded to other problems in medicine, especially those in which coincident “omics” covariates and clinical responses are simultaneously captured.

Kernel-penalized regression models for taxon-specific associations in microbiome studies

Timothy Randolph\textsuperscript{1}, Sen Zhao\textsuperscript{2} and Ali Shojaie\textsuperscript{2}

\textsuperscript{1}University of Washington
\textsuperscript{2}University of Pennsylvania School of Medicine

Principal coordinates analysis (PCoA) with respect to a phylogenetic (dis)similarities between taxon abundance vectors is one of the most common tools for visualizing data from microbiome studies. When the points in such a plot are labelled by an outcome (e.g., phenotype or class membership), these ordination plots imply a low-dimensional regression model between the outcome and abundance data. It is well established how to test this association using PERMANOVA or MrIrKAT, or to fit a model using a few latent vectors, but it is not obvious how to recover the identification of taxa that are associated with the outcome. Within our “kernel penalized regression” framework, these taxon-specific associations are estimated, even when the PCoA plot is based on a non-Euclidean dissimilarity such as UniFrac. This approach also allows one to incorporate the appropriate geometry of compositional data (relative abundances) into the structure of penalized regression. Additionally, inferential methods for this framework provide a means for assigning significance to each individual taxon for its association with the outcome.

Data exploratory methods for microbiome data analysis

Hong Gu and Toby Kenney
Dalhousie University

Recent research has shown the enormous influence of the microbiome on our lives. However analyzing the microbiome data generated from the High-throughput sequencing technique is incredibly challenging. The special characteristics of these data invalidates the application of most of the standard multivariate data exploration techniques. The most commonly used dimension reduction methods, such as principal component analysis, cannot be directly used in such data. Although it is common to apply principal coordinate analysis based on some pairwise dissimilarity measures for all pair of observations, the results are not directly interpretable according to the variables. We explore the use of a modified version of principal component analysis and non-negative matrix factorization as basic data exploration methods for finding the main structure of the microbiome data.

Session 78: Complex and Large Data Analysis

Jointly analysis of the NMR and MS data by simultaneous functional data deconvolution

Yiwen Liu, Wenyuan Zhong, Ping Ma and Arthur Edison
University of Georgia

Metabolomics provides the systematic, unbiased analysis of low-weight small molecules or metabolites. Metabolomics analysis relies on two primary techniques: mass spectrometers (MS) and nuclear magnetic resonance (NMR) spectrometers. MS is highly sensitive in ascertaining the molecular formula of a compound, but it provides very limited information about the chemical structures if no prior structure information of the compound is available. NMR, on the other hand, provides detailed information about the structure, dynamics, reaction state, and chemical environment of molecules by monitoring the activities of atomic nuclei in a metabolite. However, the wide application of NMR is hindered by the necessity of long culturing times, the need for highly trained laboratory personnel, and the requirement of expensive maintenance. Thus, a joint analysis that can borrow the strengths and overcome the shortcomings of the two techniques is highly desirable.

To utilize the NMR and MS data, we observe that the proportion of each metabolite in a sample is ???xed no matter it is measured by NMR or MS, and the proportions vary from sample to sample although the spectra of both MS and NMR remain the same. Using this fact, we propose to develop a functional data deconvolution algorithm to simultaneously quantify the MS and NMR spectra and quantify the metabolite changes across different samples. The key of this approach is to borrow knowledge from multiple samples in two platforms. The proposed method provides an analytical tool to (1) identify the chemical structure of a metabolite; (2) improve the sensitivity and specificity of metabolites detection by borrowing the strength of MS and NMR signals.
Functional Joint Model with application of Alzheimer’s disease

Kan Li and Sheng Luo

The University of Texas at Houston

Objective: We propose a functional joint model to account for functional covariates in both longitudinal submodel and survival submodel under the joint modeling framework. The parameters of this joint model are estimated in the maximum likelihood framework. The model is flexible to incorporate a number of features both in the field of joint longitudinal-survival model and in functional data analysis. Our proposed model is evaluated by simulation studies and is applied to the ADNI study, a motivating clinical trial.

Methods: We develop a functional joint model to account for functional covariates in both longitudinal submodel and survival submodel under the joint modeling framework. The parameters of this joint model are estimated in the maximum likelihood framework. The model is flexible to incorporate a number of features both in the field of joint longitudinal-survival model and in functional data analysis. Our proposed model is evaluated by simulation studies and is applied to the ADNI study, a motivating clinical trial.

Findings: The model is flexible to incorporate a number of features both in the field of joint longitudinal-survival model and in functional data analysis. Our proposed model is evaluated by simulation studies and is applied to the ADNI study, a motivating clinical trial.

Conclusion: The model is flexible to incorporate a number of features both in the field of joint longitudinal-survival model and in functional data analysis. Our proposed model is evaluated by simulation studies and is applied to the ADNI study, a motivating clinical trial.
ter than SPIEC-EASI for inferring direct microbial interactions of a mouse skin microbiome data. The gCoda is freely available from https://github.com/huayingfang/gCoda under GNU LGPL v3.

Identifiability and Inference of Causal Effects with High-Dimensional and Invalid Instruments
Changjing Wu, Minghua Deng and Wei Lin
Peking University
willin@math.pku.edu.cn
In genetical genomics studies, it is desirable to use genetic variants as instruments to estimate the causal effects of gene expressions on a clinical trait. The high dimensionality and pleiotropic effects of genetic variants bring about new challenges in choosing relevant and valid instruments. In an instrumental variable model with high-dimensional and possibly invalid instruments, we present novel identifiability conditions and propose a two-stage regularization method for identifying and estimating the causal effects. We further develop inferential procedures for obtaining debiased estimates and constructing confidence intervals. The proposed method is efficiently implemented and theoretical guarantees are provided. The usefulness of our method is demonstrated on simulated and real data.

Session 80: Quantile regression and inference for high-dimensional problems

Principal Quantile Regression for Sufficient Dimension Reduction with Heteroscedasticity
Chong Wang, Seung Jun Shin and Yichao Wu
1North Carolina State University
2Korea University
wu@stat.ncsu.edu
Sufficient dimension reduction (SDR) is a successful tool for reducing data dimensionality without stringent model assumptions. In practice, data often display heteroscedasticity which is of scientific importance in general but frequently overlooked since a primal goal of most existing statistical methods is to identify conditional mean relationship among variables. In this article, we propose a new SDR method called principal quantile regression (PQR) that efficiently tackles heteroscedasticity. PQR can naturally be extended to a nonlinear version via kernel trick. Asymptotic properties are established and an efficient solution path-based algorithm is provided. Numerical examples based on both simulated and real data demonstrate the PQR’s advantageous performance over existing SDR methods. PQR still performs very competitively even for the case without heteroscedasticity.

Variable Selection for High-Dimensional Additive Quantile Regression
Ben Sherwood and Adam Maidman
1University of Kansas
2University of Minnesota
ben.sherwood@ku.edu
Additive quantile regression is used to estimate a conditional quantile with few assumptions about the relationship between the response and the predictors. To estimate such a model in the high-dimensional setting we propose using B-splines to estimate the potentially nonlinear relationships and a nonconvex group penalty for variable selection. Estimation and variable selection properties of the method will be presented. In addition, an algorithm for approximating the nonsmooth and nonconvex objective function is presented.

Censored quantile regression in high dimensional survival data
Qi Zheng, Limin Peng and Xuming He
1University of Louisville
2Emory University
3University of Michigan
qi.zheng@louisville.edu
Censored quantile regression (CQR) has emerged as a useful regression tool for survival analysis. Stochastic integral based estimating equations are commonly used in the estimation of CQR, and pose new challenges in the analysis of CQR for high dimensional survival data. In this work, we study the high dimensional CQR simultaneously over a continuum of quantile indices. We propose a two-step penalization procedure, which accommodates stochastic integral based estimating equations and properly addresses the associated complications. We establish the uniform convergence rates for the proposed estimators and investigate the properties of weak convergence and variable selection. We conduct extensive numerical studies to confirm our theoretical findings and illustrate the practical utility of our proposals.

Session 81: Recent Advances on Statistical Modeling for Cancer Genomics Data

Absolute Mutation-Specific Risk of Breast and Ovarian Cancer in BRCA1 or BRCA2 Mutation Carriers
Jinbo Chen, Xinglei Chai and Timothy Rebbeck
1University of Pennsylvania Perelman School of Med
Dana Farber Cancer Institute
jinboche@mail.med.upenn.edu
Breast and ovarian cancer risk for women carrying BRCA1/2 mutations varies by mutation location and type. To date, mutation-specific absolute risk has not been available for use in clinical risk assessment and decision-making. We calculate the absolute risk of breast and ovarian cancer due to BRCA1/2 mutations occurring in the breast (BCCR) and ovarian (OCCR) cancer cluster regions. Accounting for use of risk-reducing salpingo oophorectomy, we generated baseline risks and used BCCR- and OCCR-specific hazard ratio estimates to generate mutation-specific absolute cancer risks. Our results showed that absolute risk estimates for BRCA1 and BRCA2 mutation carriers vary widely with mutation locations, and may inform cancer risk assessment and decision-making in BRCA1/2 mutation carriers. Specifically, Variation in breast and ovarian cancer absolute risk was strongly associated with occurrence of mutations in the BCCR and OCCR regions. Breast cancer risks for carriers of BCCR mutations were as high as 33% greater in BRCA1 and 42% greater in BRCA2 compared with the lowest breast cancer risks in women with OCCR mutations. Ovarian cancers risks for carriers of OCCR mutations were as high as 37% greater in BRCA1 and 136% greater in BRCA2 compared with the lowest ovarian cancer risks in women with BCCR mutations.

The MiAge Calculator: a DNA methylation-based mitotic clock calculator of Human tissue types
Ahrim Youn and Shuang Wang
1The Jackson Laboratory for Genomic Medicine
2Columbia University
sw2206@columbia.edu
Cell division is known to be important in cancer. To estimate the number of cell divisions (mitotic age) of a given tissue type between individuals is of great interest as that allows their stratification of prospective cancer risk. Here we introduce the MiAge Calculator, a
mitotic clock calculator based on a novel statistical framework, the MiAge model. MiAge is designed to quantitatively estimate mitotic age of a tissue using the stochastic replication errors accumulated in the epigenetic inheritance process during cell divisions. With MiAge, the MiAge Calculator was built using the training data of eight TCGA cancers. We tested the MiAge Calculator using the testing data of five other TCGA cancers. We further showed that the estimated mitotic clock of the MiAge Calculator is universally accelerated in tumor tissues than in adjacent normal tissues within each of the thirteen cancer types studied. Across the thirteen cancer types, we showed that worse cancer survivals are associated with more accelerated mitotic clock in tumor tissues. Importantly, we demonstrated the utility of mitotic clock by showing that the integration of mitotic clock and clinical information leads to an improved survival prediction in six out of the thirteen cancers studied.

Identifying the effects of multiple dichotomous mediators on a dichotomous outcome

♦ Stephannie Shih¹, Yen-Tsung Huang² and Simin Liu¹

¹Brown University
²Brown University; Academia Sinica

Risk factors of complex diseases such as cancer may occur at different stages in life and act as intermediary effects between an exposure and disease occurrence. While the existing methodology in multi-mediator models focused on continuous mediators, the intermediary events such as occurrence of pre-disease health event may be dichotomous. In order to assess the causal pathway in this type of problem, we developed a multi-mediator model that can quantify the effect of having an exposure of interest on disease risk through two dichotomous mediators. Our method provides closed form formula for the mediation effects through different paths, and does not require rare outcome assumption. Variance of the effects can be estimated using delta method or bootstrapping. The proposed method under finite sample was evaluated via simulation studies. We then applied our method to data from the Women’s Health Initiative to assess the causal pathway of type 2 diabetes.

On the use of validation method for molecular classification

Li-Xuan Qin, Huei-Chung Huang and Colin Begg
Memorial Sloan Kettering Cancer Center

qinl@mskcc.org

Reproducibility of scientific experimentation has become a major concern, due to the perception that many published biomedical studies cannot be replicated. In this talk we draw attention to the connection between inflated over-optimistic findings and the use of cross-validation for error estimation in molecular classification studies, in the presence of confounding handling effects in the data. We demonstrate this important yet overlooked complication of cross validation using a unique pair of datasets on the same set of tumor samples. One dataset was collected with uniform handling to prevent handling effects; the other dataset was collected without uniform handling and exhibited handling effects. The paired datasets were used to estimate the biological effects of the samples and the handling effects of the arrays in the latter dataset, which were then used to simulate data using virtual re-hybridization following various array-to-sample assignment schemes. Our study showed that (1) cross-validation tended to underestimate the error rate when the data possessed confounding handling effects, (2) depending on the relative amount of handling effects, normalization may further worsen the under-estimation of the error rate, (3) balanced assignment of arrays to comparison groups allowed cross-validation to provide an unbiased error estimate. Our study demonstrates the benefits of balanced array assignment for reproducible molecular classification and calls for caution on the routine use of data normalization and cross-validation in such analysis.

Session 82: Statistical considerations in Multi-regional trials

Some Recent Advances on Statistical Approaches for Planning Multi-Regional Clinical Trials

H.M. James Huang, Xiangmin Zhang and Sue-Jane Wang
Food and Drug Administration

xianmin.zhang@fda.hhs.gov

Multiregional clinical trials (MRCT) are increasingly employed in the development of drug products. One issue that the MRCT faces is the interpretability of the trial results at the presence of heterogeneous regional treatment effects. The regional heterogeneity can be modeled through random effects and different random effects models have been proposed. On the other hand, if one believes that the regions are heterogeneous, the regional heterogeneity needs to be accounted for when planning the sample size at the trial design stage. In this talk, we discuss the type I error and power under different statistical models used to test the global treatment effect for the MRCT. We show that ignoring the regional heterogeneity could lead to a severe inflation of type I error and decrease of power. We assess the probability that some regions show false negative effects by chance. We then introduce an alternative method for sample size planning to account for regional heterogeneity in the MRCT.

Multi-regional Biosimilarity Studies

♦ Yu-Wei Chang¹ and Xia Qi²

¹Boehringer Ingelheim
²Temple University

changvick@gmail.com

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.

Design and Evaluation of Multiregional Clinical Trials with Heterogeneous Variability across Regions

♦ Chin-Fu Hsiao and Chiieh Chiang

National Health Research Institutes

chinfu@nhri.org.tw

To speed up drug development to allow faster access to medicines for patients globally, conducting multi-regional trials incorporating subjects from many countries around the world under the same protocol may be desired. Several statistical methods have been pro-
posed for the design and evaluation of multiregional clinical trials (MRCTs). Most of these approaches, however, assume a common variability of the primary endpoint across regions. In practice, this assumption may not be true due to differences across regions (e.g., differences in ethnic factors and/or medical culture/practice). In this talk, we concern the case that the mean of response variable is fixed but the variances across regions are different. We evaluate the drug efficacy in overall region when un-equal variability of the primary endpoint across regions is assumed in the MRCT. A method for sample size determination of the MRCT is also proposed. As we know, the aim of a MRCT is to show the efficacy of a drug in various global regions, and concurrently to evaluate the possibility of applying the overall trial results to each region. Therefore, we also address consideration on the determination of the number of subjects in each region to establish the consistency of treatment effects across regions.

Session 83: Combining Endpoints in Clinical Trials

Weighted win loss approach for analyzing prioritized outcomes
Xiaodong Luo1, Junshan Qiu2, Steven Bai2 and Hong Tian2
1Sanofi US
2US Food and Drug Administration
3Janssen Research and Development

To analyze prioritized outcomes, Buyse (2010) and Pocock et al. (2012) proposed the win loss approach. This approach uses a layered comparison procedure to account for the order of priorities of different outcomes therefore has attracted a lot of interests in methodological development and real-life application. In this talk, I will first study the relationship between the win loss approach and the traditional survival analysis on the time to the first event. We then propose the weighted win loss statistics to improve the efficiency of the un-weighted methods. Contribution index is used to supplement the win loss approach for a better interpretation of the results. Simulation studies and real data analysis demonstrated the characteristics of the proposed statistics.

Generalized pairwise comparisons for prioritized outcome measures
Marc Buyse1, Julien Peron2, Joris Giai3 and Pascal Roy3
1IDDI Inc., San Francisco
2Medical Oncology Department, CH Lyon Sud, Lyon
3Hospices Civils de Lyon, Service de Biostatistique

Generalized pairwise comparisons extend the Mann-Whitney form of Wilcoxon's test, or Gehan’s generalization of this test for two samples of (possibly censored) observation. The test uses all pairwise comparisons between two patients, one in the treatment arm and one in the control arm, in terms of one or several prioritized outcomes. Each pair favors treatment, control, or neither. The treatment benefit is the difference between the proportion of pairs in favor of treatment less the proportion of pairs in favor of control; this statistic is called the “net chance of a better outcome”. For a single variable, this measure of treatment benefit has a simple relationship with traditional measures of benefit such as the risk difference (for binary outcomes), the mean difference (for continuous outcomes), or the hazard ratio (for right-censored outcomes). The pairwise comparison approach easily incorporates a threshold of clinical relevance in the analysis. For instance, if survival is the outcome of interest, a threshold of m months can be defined to estimate the net chance of a longer survival by at least m months. The pairwise comparison approach also allows several prioritized outcomes, for instance time to tumor progression and overall survival, to be analyzed simultaneously.


Measures of Clinical Benefit in Immuno-Oncology Studies
Luping Zhao, Yifan Huang, Yiduo Zhang, Kelvin Shi, Wenmei Huang and Pralay Mukhopadhyay
AstraZeneca

Understanding the value of a new therapeutic agent in treatment of cancer is important to both caregivers and patients. The recently developed ASCO and ESMO value framework attempts to address this complex problem using a multi-dimensional approach that evaluates benefit in terms of efficacy, safety and QoL. Here we compare their recommendations on measuring clinical benefit in terms of efficacy, for different classes of anti-cancer agents (immunotherapy, chemotherapy, small molecules). In particular, we focus on the emerging immuno-oncology agents where there is a possible underlying delayed treatment effect. We compare the performance of commonly used statistics (hazard ratio, median, restricted mean survival and survival at landmark time points) with the composite scores proposed by ASCO and ESMO-MCBS for clinical benefit assessment under different treatment scenarios (PH, delayed treatment effect, belly shaped KM curves). A novel method to measure clinical benefit is proposed using a weighted combination of these standard measures. The existing statistics/scores and the proposed method are evaluated using both simulated and real clinical trial data.

Analyzing Multiple Endpoints Simultaneously in Randomized Clinical Trials using GPC
Joris Giai1, Julien Peron1, Delphine Mauvrot-Bouch1, Pascal Roy3 and Marc Buyse2
1Hospices Civils de Lyon / CNRS UMR 5558 / UCBL
2International Drug Development Institute (IDDI)

Background: In randomized clinical trials, efficacy and safety outcomes are usually analyzed and reported independently. We sought to apply Generalized Pairwise Comparisons (GPC) as an innovative method of assessing the benefit-risk balance using data from three completed randomized controlled trials. Methods: In both PRODIGE 4 and NCIC CTG PA.3 trials in pancreatic cancer, the first prioritized outcome was overall survival time with a two months clinically relevant threshold. The second prioritized outcome was toxicity. In the PROFIL N-of-one trial testing Sildenafil in Raynaud’s Phenomenon, first to third prioritized outcomes were daily crises number with a single crisis threshold, Raynaud’s Condition Score with a two points threshold and daily sum of crises dura-
tion with a ten minutes threshold respectively. Sensitivity analyses were performed. Results: In PRODIGE 4 trial, the net chance of a better outcome favored significantly the FOLFIRINOX group (24.7; p=0.001). suggesting a favorable benefit-risk balance of FOLFIRINOX versus Gemcitabine. In NCIC CTG PA.3, the method indicated no statistically significant overall treatment effect in favor of Erlotinib added to Gemcitabine versus placebo (-3.6; p=0.51). In PROFIL, the net chance of a better outcome favored significantly Sildenafil over placebo after pooling patients’ results from the stratified analysis (5.6; p=0.04). Those results were robust throughout most sensitivity analyses. Conclusions: Generalized pairwise comparisons are useful to perform a quantitative assessment of the benefit-risk balance of a new treatment as compared with a standard therapy. It provides a clinically intuitive way of comparing patients with respect to all important efficacy and toxicity outcomes, with full flexibility as to the priority of each outcome, and a threshold of clinical significance.

Session 84: Biomarker development: Ideas and practicalities

Biomarker Development: Ideals and Practicalities

Karla Ballman
Weill Cornell Medicine
Kab2053@med.cornell.edu

The Holy Grail in medicine, and especially in cancer treatment, is to determine which patients require more aggressive treatment (prognostic biomarkers) and to select the treatment that would result in the greatest benefit for the patient (predictive biomarkers). The ideal study design and analysis depends on the intended use of the biomarker. Ideal study designs and analyses for the different types of biomarkers will be reviewed. I will then discuss the types of studies that are often done in practice due to the patient cohorts and data that are available. For these scenarios, I will indicate the types of conclusions that can be supported by the data. Examples from published studies will be used to illustrate key points.

Designing the trial of anti-PD-I therapy in tumors with mismatch repair deficiency: Innovative study

Hao Wang
Johns Hopkins University School of Medicine
hwang76@jhmi.edu

Advances in understanding of cancer biology and biomarker discoveries have led to exciting development of targeted therapies and immunotherapy. Unlike the traditional clinical trials that are designed to answer a single efficacy question, the trials for these types of therapy must evolve to integrate science and identify predictive biomarkers. In this talk I present a design strategy for a study of a genetic biomarker to select patients for immunotherapy. The adaptive parallel-cohort and basket designs for separate components of the study allows to confirm a predictive biomarker that predicts response to immunotherapy, and furthermore, demonstrate that immunotherapy based solely on a genomic marker is efficacious irrespective of tumor histology. These novel design schemes enable the trial to meet its multifaceted research objectives with a reduction in cost, resource utilization.

Biomarker development: Ideas and practicalities

Bruce Trock
Johns Hopkins University
btrck@jhmi.edu

Rapid advances in omics technology, understanding of molecular mechanisms, and methods to analyze high dimensional data has led to a huge increase in studies of cancer prognostic biomarkers, signatures and models. This has been accompanied by increased interest in methods to evaluate how well a new biomarker or signature can improve the ability to predict adverse outcomes such as metastasis or death. Performance evaluation most commonly involves comparison of nested models with respect to the area under the receiver operating characteristic curve (AUC) or the concordance index for survival outcomes, the net reclassification index (NRI), and decision curve analysis (DCA). All of these methods have advantages and shortcomings. A problem common to all is how to translate the measure of performance improvement into clinically meaningful terms that physician and patient can understand, and which can guide a treatment decision. This problem is particularly difficult in prostate cancer, because it frequently has a very non-aggressive course, has a long natural history, a large number of treatment options, and clinical and pathology prognostic factors that already predict outcomes with a high degree of accuracy. This talk will examine problems in evaluating prognostic biomarkers in prostate cancer, and consider a potential new performance measure based on a modification of reclassification statistics.

Session 85: Statistical Analysis for Non-normal Data

High Dimensional Variable Selection for Censored Quantile Regression

Naveen Naidu Narisetty\(^1\) and Xuming He\(^2\)

\(^1\)University of Illinois at Urbana Champaign
\(^2\)University of Michigan

naveen@illinois.edu

Quantile regression provides a more comprehensive relationship between a response and covariates of interest as compared to mean regression. Quantile models are particularly advantageous for censored data, which are commonly encountered in many biological and economic applications. When the response variable is subject to censoring, estimating conditional mean requires strong distributional assumptions on the error whereas conditional quantiles can be estimated distribution-free. Although conceptually appealing, quantile regression for censored data is challenging due to both computational and theoretical difficulties arising from non-convexity of the objective function involved. This is even more prominent when the covariates are high-dimensional and can be more in number than samples. We consider a Powell’s objective function based likelihood and place appropriate priors on the regression parameters in a Bayesian framework. In spite of the non-convexity of Powell’s objective function and mis-specification of the likelihood, we show that the posterior distribution is strong selection consistent even when the number of variables is much larger than the sample size. The posterior can be sampled from a Gibbs sampling algorithm that has complexity linear in number of variables, avoiding the computational difficulties associated with optimization of a non-convex objective function.

Measures of Income Inequality Focusing on the Lower and Middle Classes

Joseph Gastwirth
Department of Statistics, George Washington Univ.
jigast@gwu.edu

The substantial increase in economic inequality in favor of the upper income group in the United States other nations opiningations during the past 30 years has become a major public policy concern.
Under-dispersion models: models that are "under the radar"

Kimberly Sellers1 and Darcy Morris2
1Georgetown University, and U.S. Census Bureau
2U.S. Census Bureau
kfs7@georgetown.edu

The Poisson distribution is a benchmark for modeling count data. Its equi-dispersion constraint, however, does not accurately represent real data. While most datasets express over-dispersion, more examples are surfacing that display under-dispersion, warranting the need to study this phenomenon more closely and propose models that can better describe such data structures. This talk (based on the Sellers and Morris (2017) work) highlights several distributions that can model under-dispersed data, and compares their performance on an applied dataset.


A class of semiparametric transformation models and their applications

Guoqing Diao
George Mason University
jdiao@gmu.edu

Non-normally distributed data occur frequently in many applications. A useful class of models for handling non-normal data are the semiparametric transformation models. Under these models, after an unknown strictly increasing transformation, the outcome variable is assumed to be linearly related to the covariates. This class of models include the commonly-used survival models such as the Cox proportional hazards model and the proportional odds model as special cases. In this research, we provide a review of the semiparametric transformation models. Specifically, we discuss likelihood-based estimation and inference procedures for the unknown parameters, as well as the asymptotic properties of the nonparametric maximum likelihood estimators. To demonstrate the flexibility and usefulness of the semiparametric transformation models, we discuss their applications in various fields including survival analysis, statistical genetics, diagnostic medicine, and graphics.

Session 86: Statistical methods in oncology trials

Predicting Data Cut-off Date in Event-Driven Clinical Trials When There is a Lag in Adjudication

Jiawang Wang and Lei Fu
Celgene Corporation
jiawang@celgene.com

The timings of interim/final analysis for event-driven clinical trials are typically determined by the accrual of the events of interest during the study. In many situations, especially for confirmatory Phase III clinical, both investigator reported events (IRE) and central adjudicated events (CAE) are collected, with the timings of the primary analyses driven by the number of CAE. Predicting analysis time based on date from CAE alone could lose accurate because data from CAE is usually associated with a lag for a variety of reasons. However, the corresponding data from IRE is usually more updated. In this article, we propose a Bayesian method to predicting the analysis time based on the predicted number of CAE by modeling the joint distribution of CAE and IRE, with the intention to borrow strength from the correlated and more updated information from IRE. In our proposed method, CAE and IRE are jointly modeled with a Marshall-Olkin bivariate exponential or Weibull distribution. Some simulations are performed to evaluate the performance of the proposed approach and to compare it with method that use CAE information alone. An application to an ongoing clinical trial is also described.

Correcting Treatment Effect for Treatment Switching in Randomized Oncology Trials with a Generalize

Jin Zhang
Merck & Co.
jin.zhang3@merck.com

In randomized oncology trials, patients in the control arm are sometimes permitted to switch to receive experimental drug after disease progression. This is mainly due to ethical reasons or to reduce the patient dropout rate. While progression-free survival is not usually impacted by crossover, the treatment effect on overall survival can be highly confounded. The Rank-Preserving Structural Failure Time (RPSFT) model is one of the main randomization-based methods used to adjust for confounding in the analysis of overall survival. While the RPSFT has been extensively studied, one major assumption made is that the treatment effect is constant regardless of when treatment switching occurs. In practice, this assumption is difficult to verify and sometimes not valid. In this manuscript, we generalized the RPSFT to accommodate the scenario where the treatment effect changes after switching. We compared the RPSFT and the generalized RPSFT via extensive simulations and then walked through the analysis using the generalized RPSFT in a real clinical trial.

The Utility of the Yang-Prentice Model when the Cox Proportional Hazards Assumption is Violated

Carl Di Casoli
Celgene Corporation
cdicasoli@celgene.com

In many immuno-oncology studies, there could be a possibly delayed treatment effect and hence, the situation where the survival curves between treatment and placebo arms can cross. In these situations, the standard Cox proportional hazards model may not be appropriate. Instead, we examine the utility of the Yang-Prentice (2005) model, a two-sample model that can take into account the potential crossing of survival curves. This model involves two parameters that represent both short-term and long-term hazard ratios and is a semiparametric model with a completely unspecified distribution at baseline, nesting both the Cox proportional hazards model and proportional odds model as special cases. Yang and Prentice implemented a pseudo maximum likelihood estimation approach which is now present in the recently developed R packages "ELYP" and "YPmodel" and can easily be implemented. In addition to the pseudo maximum likelihood approach, this author has examined several Bayesian methods applied to the Yang-Prentice model. In this talk, several real data examples will demonstrate the utility of
the various approaches.

Session 87: Mixed-effects models in genomics, health and psychology

A unified powerful set-based test for sequencing data analysis of GxE interactions

Yu-Ru Su, Chong-Zhi Di and Li Hsu
FRED HUTCHINSON CANCER RESEARCH CENTER

ysu@fredhutch.org

The development of next generation sequencing technologies has allowed researchers to study comprehensively the contribution of genetic variation particularly rare variants to complex diseases. To date many sequencing analyses of rare variants have focused on marginal genetic effects and have not explored the potential role environmental factors play in modifying genetic risk. Analysis of GxE (gene-environment interaction) for rare variants poses considerable challenges because of variant rarity and paucity of subjects who carry the variants while being exposed. To tackle this challenge, we propose a hierarchical model to jointly assess the GxE effects of a set of rare variants e.g., in a gene or regulatory region, leveraging the information across the variants. Under this model, GxE is modeled by two components. The first component incorporates variant functional information as weights to calculate the weighted burden of variant alleles across variants, and then assess their GxE interaction with the environmental factor. Since this information is a priori known, this component is fixed effects in the model. The second component involves residual GxE effects that have not been accounted for by the fixed effects. In this component, the residual GxE effects are postulated to follow an unspecified distribution with mean 0 and variance $\tau^2$. We develop a novel testing procedure by deriving two independent score statistics for the fixed effects and the variance component separately. We propose two data-adaptive combination approaches for combining these two score statistics and establish the asymptotic distributions. An extensive simulation study shows that the proposed approaches maintain the correct type I error and the power is comparable to or better than existing methods under a wide range of scenarios. Finally we illustrate the proposed methods by a exome-wide GxE analysis with NSAIDs use in colorectal cancer.

Modeling of Between- and Within-Subject Variances Using Mixed Effects Location Scale Models

Donald Hedeker$^1$ and Robin Mermelstein$^2$

$^1$University of Chicago
$^2$University of Illinois at Chicago
hedeker@uchicago.edu

Intensive longitudinal data are increasingly encountered in many research areas. For example, ecological momentary assessment (EMA) and/or mobile health (mHealth) methods are often used to study subjective experiences within changing environmental contexts. In these studies, up to 30 or 40 observations are usually obtained for each subject over a period of a week or so, allowing one to characterize a subject’s mean and variance and specify models for both. In this presentation, we focus on an adolescent smoking study using EMA where interest is on characterizing changes in mood variation. We describe how covariates can influence the mood variances and also extend the statistical model by adding a subject-level random effect to the within-subject variance specification. This permits subjects to have influence on the mean, or location, and variability, or (square of the) scale, of their mood responses. These mixed-effects location scale models have useful applications in many research areas where interest centers on the joint modeling of the mean and variance structure.

On estimating within-person relations using mixed-effects modeling

Jijuan Wang$^1$, Qian Zhang$^2$, Scott Maxwell$^3$ and Cindy Bergeman$^1$

$^1$University of Notre Dame
$^2$Florida State University
$^3$University of Nebraska, Lincoln

This study further extends current discussion of how to disaggregate and estimate between-person and within-person relations with longitudinal data using multilevel modeling. In the literature, person-mean centering has been recommended and used. The goal of this study is to evaluate whether and when person-mean centering can accurately estimate fixed-effects within-person relations using multilevel modeling. We analytically derived that when individual differences exist in intrapersonal standard deviations, the estimated fixed-effects “within-person” relations from person-mean centering coupled with global standardization (denoted as traditional person-mean centering) may be asymptotically biased, even when there are no trends in either of the time-varying predictor or outcome variables. Specifically, we found that interindividual differences in intrapersonal/person standard deviations play an important role in fixed-effects within-person relations, but are not appropriately considered/modelled with the traditional person-mean centering approach. To resolve the problem, we recommend person-mean-SD standardization for appropriately disaggregating within- and between-person relations and accurately estimating fixed-effects within-person relations.

Modeling non-normal distributions in mixed-effects and multilevel models

Zhiyong Zhang

University of Notre Dame
zhang4@nd.edu

Mixed-effects and multilevel models are widely used in social and behavioral sciences. However, typical mixed-effects models often assume that the errors are normally distributed although non-normal data may be even more common than normal data. In order to avoid possible statistical inference problems in blindly assuming normality, a general Bayesian framework is proposed to flexibly model normal and non-normal data through the explicit specification of the error distributions. Specifically, we show how model the errors using $t$, skew normal and exponential power distributions that cover a wide variety of potentially observed data. A simulation study shows when the distribution of the error is correctly specified, one can avoid the loss in the efficiency of standard error estimates. A real example on the analysis of mathematical ability growth data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 is used to show the application of the proposed methods.

Session 88: Making Sense of Big Omics Data “C Recent Advances in Statistical and Computational Methods

Statistical methods for Hi-C data to study DNA structure

Yun Li$^1$, Ming Hu$^2$, Joshua Martin$^1$, Zheng Xu$^3$, Bing Ren$^4$ and Karen Mohlke$^1$

$^1$University of North Carolina, Chapel Hill
$^2$Cleveland Clinic Foundation,
$^3$University of Nebraska, Lincoln,
Chromosome conformation capture (3C) derived technologies have become increasingly popular to study the three dimensional (3D) structure of our genome and DNA looping regulatory interactions. However, methods to analyze data from 3C-derived technologies are truly in their infancy. We (Xu et al 2015 PMID: 26543175) present a hidden Markov random field (HMRF) based Bayesian framework to detect long range chromatin interactions, accounting for spatial dependency, empirically non-negligible but been ignored by existing methods. We further proposed a computationally as well as statistically more efficient method (Xu et al 2016 PMID: 27851967), elucidating new insights gained from DNA structure, integrated with transcriptional, regulatory, genomewide association results. Lastly, we have developed a visualization tool (http://yunliweb.its.unc.edu/hugin/) to facilitate efficient and effective data mining across a compendium of chromatin interaction data, and to suggest potential target genes of GWAS variants in relevant tissues.

A Dirichlet mixture model for clustering droplet-based single cell transcriptomic data
Zhe Sun¹, Ting Wang¹, Ming Hu² and Wei Chen¹
¹University of Pittsburgh
²Cleveland Clinic Foundation
weichen.mich@gmail.com

Single cell transcriptome sequencing (scRNA-seq) has become a revolutionary tool to study transcriptomic profiles of individual cells. The recently developed droplet-based platform enables efficient parallel processing of thousands of cells with direct counting of transcript copies using Unique Molecular Identifier (UMI). Despite the technology advance, statistical methods and computational tools are still lacking for analyzing droplet-based single cell transcriptomic data. We developed a Dirichlet mixture model for clustering droplet-based Single Cell transcriptomic data. This approach explicitly models UMI count data from scRNA-seq experiments and characterizes variations across different cell clusters via Dirichlet mixture prior. We performed comprehensive simulations and compared the performance with K-means, CellTree, and Seurat. In addition, we analyzed public and in-house scRNA datasets from multiple studies of complex diseases. We demonstrate that our method has a substantially improved clustering accuracy and stability and provides clustering uncertainty for downstream analysis and better biological interpretations. The program has been implemented in a user-friendly R package.

Bayesian hierarchical model with multi-layer overlapping group structure in omics applications
Li Zhu and George Tseng
Department of Biostatistics, Univ. of Pittsburgh
ctseng@pitt.edu

Rapid advances in high-throughput genomics experiments, such as microarray and next-generation sequencing, have increased availability of multi-level omics data (e.g. mRNA expression, miRNA expression, methylation, etc.). Integration of multi-level omics data for biomarker association and outcome prediction has brought new statistical challenges. One great challenge is to incorporate prior biological knowledge of multi-layer overlapping group structure in the modeling. The hierarchical structure of Bayesian models makes it a flexible and convenient modeling choice. In this paper, we propose a Bayesian binary masking model with effective feature and group selection in each layer for this purpose. Simulation and application to TCGA breast cancer data show more accurate feature selection and prediction, compared to existing group lasso or counterparts of Bayesian models. We also demonstrate consistency and asymptotic normality of feature selection and parameter estimates using the posterior median estimate from our model, under mild conditions. The result shows improved capacity of incorporating complex prior knowledge structure for better prediction accuracy and better feature (gene) selection to understand disease and enhance precision medicine.

A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome
Anthony Schmitt¹, Ming Hu² and Bing Ren¹
¹University of California San Diego
²Cleveland Clinic
afhuming@gmail.com

The three-dimensional configuration of DNA is integral to all nuclear processes in eukaryotes, yet our knowledge of the chromosome architecture is still limited. Genome-wide chromosome conformation capture studies have uncovered features of chromatin organization in cultured cells, but genome architecture in human tissues has yet to be explored. Here, we report the most comprehensive survey to date of chromatin organization in human tissues. Through integrative analysis of chromatin contact maps in 21 primary human tissues and cell types, we found topologically associating domains highly conserved in different tissues. We also discover genomic regions that exhibit unusually high levels of local chromatin interactions. These frequently interacting regions (FIREs) are enriched for super-enhancers and are near tissue specifically expressed genes. They display strong tissue-specificity in local chromatin interactions. Additionally, FIRE formation is partially dependent on CTCF and the Cohesin complex. We further show that FIREs can help annotate the function of non-coding sequence variants.

Session 89: Statistical Applications in Population Health Sciences

Using genetic markers in risk prediction and pharmacogenomics studies for lung cancer
Chao Hsiung
National Health Research Institutes
hsiung@nhri.org.tw

(I)Risk Prediction study: Epidemiology studies have shown that the incidence of lung cancer in never-smoking women is particularly high in Asia. Several SNPs (single nucleotide polymorphisms) have been identified to be associated with lung cancer susceptibility among never smokers by genome-wide association studies (GWAS). We constructed a multivariate model for the estimate of absolute risk of a female developing lung cancer within a 6-year period, based on age, conventional epidemiological risk factors and these known susceptibility SNPs. This model is potentially useful to identify high-risk individuals in designing lung cancer screening programs.

(II)Pharmacogenomics study: Patients with non-small cell lung cancer (NSCLC) with mutated epidermal growth factor receptor (EGFR) are relatively sensitive to EGFR-tyrosine kinase inhibitor (TKI) treatment and have longer progression-free survival (PFS) when treated with EGFR-TKI compared with platinum-based chemotherapy. However, many patients with advanced NSCLC who
have mutated EGFR do not respond to first-line EGFR-TKI treatment and still have shorter PFS. The aim of this study was to identify genetic variants associated with PFS among patients with lung adenocarcinoma who were treated with first-line EGFR-TKIs. We identified SNPs at 4q12 associated with PFS at genome-wide significance and with an estimated hazard ratio of more than 4. This association was also replicated in a larger but similar cohort and in an independent NSCLC cohort. Follow-up functional analyses showed that these SNPs were associated with the expression of EGFR, which encodes the TKI target, and with a nearby gene neuromedin-U, which encodes a G protein’ coupled receptor ligand known to be involved in the progression of NSCLC. Genetic variants in 4q12 merit further investigation to assess their potential as pharmacogenomic predictors for and to understand the biology underlying its influence on PFS in patients treated with TKI therapy.

**Longitudinal studies of children and adolescent behaviors**

**Hsing-Yi Chang**
National Health Research Institutes

Most of the adulthood behavior or health status are shaped in childhood. Thus, many researchers started to gather information of children. Some even follow the children for long time. The British Centre for Longitudinal Studies (CLS) (http://www.cls.ioe.ac.uk) documented many of their children cohorts, initiated in 1958 and 2000. The longest running cohort is also in UK. It was established by the National Survey for Health and Development (NSHD) in 1946. Their members turned 70 in 2016. The Grant Study was conducted in Harvard starting in 1939-1944 on college students, then included more disadvantaged nondelinquent inner-city youths as second cohort in 1940-1945, known as the Gleuck Study. Having gathered these valuable information, many questions can be answered. At the same time, analyzing these data is full of challenges. This talk will present the challenges basing the experiences of analyzing the children and adolescent behavior in long-term evolution (CABLE, http://cable.nhri.org.tw). The CABLE study was initiated in 2001. Samples of the 1st graders (n=2586) and the 4th graders (n=2667) were selected in Northern Taiwan, and interviewed annually till they graduated from colleges. The study was designed based on an ecological model, which assumed a person’s health and lifestyles are influenced by the individual, family, school, and community. The problems in lost to follow-up (missing), modeling the trajectories, the transition, and the patterns, as well as the factors associated with the above mention items will be discussed.

**Pharmacovigilance studies of psychotropic agents’ opportunities and pitfalls of big data**

**Hui-Ju Tsai¹, Chi-Shin Wu², Yu-Ting Tsai¹ and Ya-Wen Huang¹**

¹National Health Research Institutes
²National Taiwan University Hospital

Psychotropic agents have been widely administered in treating subjects with mental illness such as bipolar disorder, depressive disorder, delirium and schizophrenia, etc. over the past decades. It has been well recognized that psychotropic agents are the most effective agents used in clinical psychiatric practice. However, in recent years, both epidemiological and clinical studies have documented that use of psychotropic agents increases risk of developing various health outcomes such as accidental events, bone related events, gastrointestinal bleeding, cardiovascular diseases, cerebrovascular diseases or peripheral diseases. As a result, previous reports including ours have drawn substantial degree of attention with respect to drug safety of psychotropic agents, especially, among populations at high risk.

In this presentation, I will present some of my previous research results related to examining effects of psychotropic agents on cerebrovascular diseases and peripheral diseases. Three research topics will be discussed in this presentation: 1) a case-crossover study design for assessing association between the receptor-binding profiles of antipsychotics and stroke; 2) a case-control study design for investigating association between antidepressants and venous thrombembolism; 3) a case-crossover study design for examining effects of antipsychotics on the risk of ventricular arrhythmia and/or sudden cardiac death.

In sum, previous works have provided suggestive evidence that use of psychotropic agents increase risk of cerebrovascular diseases and/or peripheral diseases. Subjects taking psychotropic agents should carefully monitor potential adverse effects of psychotropic agents, especially, in the beginning of initial prescribing psychotropic agents.

**Methods and Theory for Spatial Change Set Analysis**

**Pei-Sheng Lin**
National Health Research Institutes

In many scientific studies, it is of interest to group spatial units on a lattice with similar characteristics within a group but with distinction among groups. Here we develop a novel change-set method for this purpose, as a substantive extension of the existing change-point analysis for one-dimensional data. Our method addresses unique challenges resulting from the two-dimensional space such as changes that occur abruptly in space and change sets of irregular shapes. In particular, we propose a homogeneity measure that enables the identification of change sets. We also establish quasi-likelihood estimation that accounts for covariates via change-set complementary models and spatial dependence via a working covariance. Asymptotic properties of our method are established under suitable regularity assumptions including a mixing condition for random fields. An inequality for the homogeneity measure is derived, by which we further establish the change-set selection consistency. In addition, a statistic for testing a candidate change set is shown to be consistent and asymptotically normal. Finite-sample properties are investigated in a simulation study and for illustration, our method is applied to analyze a county-based socio-economic data set.

**Session 90: New Advances on Nonparametric Methods for High-dimensional Data**

**Tree-based model for longitudinal data**

**Peng Wang and Brittany Green**
University of Cincinnati

Classification and regression tree has been broadly applied due to its simplicity of explanation, automatic variable selection, visualization and interpretation. Previous algorithms for constructing regression and classification tree models for longitudinal data suffer from the computational difficulties in estimation of covariance matrix at each node. In this paper, we proposed a regression and classification tree for longitudinal data, utilizing the quadratic inference function (QIF). Following CART approach and taking the correlation of longitudinal data into consideration, we developed a new criterion, named RSSQ, to select the best splits. The proposed approach could
incorporate the correlation between the repeated measurements on the same subject without estimation of the correlation parameter. Therefore we could improve the efficiency of the partition results and prediction accuracy. Simulation studies and real data examples are provided to illustrate the promise of the proposed approach.

**Partially Linear Single-Index Models in Ultra-high Dimension**

*Shan Yu, Lily Wang and Dan Nettleton*

**Iowa State University**

syu@iastate.edu

High-dimensional data occur very frequently in a large variety of research areas. The partially linear single-index model (PLSIM) is flexible and powerful tool for modeling this type data. In this article, we consider PLSIMs for analyzing ultra-high-dimensional data where both the number of linear components and single-index components can be much larger than the sample size. The single-index link function is approximated using polynomial spline basis functions. A fast and efficient iterative algorithm is proposed to select nonzero linear and single-index components based on different regularizations. The proposed method selects the correct model with probability approaching one under regularity conditions. The performance of the method is evaluated by simulation studies. The proposed method is also applied to a dataset on the Shoot Apical Meristem (SAM) of maize genotypes.

**Robust high-dimensional regression and variable selection**

*Yichen Qin*, Shaobo Li1, Yang Li2 and Yan Yu3

1University of Cincinnati
dqiyn@ucmail.uc.edu

We introduce a class of robust mean regression estimators for variable selection. We first explain the motivations for the key ingredient, the robustified maximum likelihood estimation. We show that it can be regarded as a mixture of minimum Kullback-Leibler distance estimation and minimum L2 distance estimation, where a robustness tuning parameter balances the trade-off of efficiency and robustness. We then propose a penalized robustified likelihood estimation for variable selection. Under appropriate conditions, we show that the proposed method is root-n-consistent, enjoys the oracle property, and has desirable robustness properties. We also discuss how to select the tuning parameter that controls the robustness. Furthermore, we consider the proposed method under a high-dimensional setting when the number of variables d can grow exponentially with the sample size n, and establish its consistency. We show that for any positive tuning parameter, the proposed method can achieve the optimal rate of convergence in the order of ln(d)/n. Finally, simulation studies and real data analysis demonstrate the advantages of the proposed method.

**Fast Covariance Estimation for Sparse Functional Data**

Luo Xiao1, Cai Li2, William Checkley2 and Ciprian Crainiceanu2

1North Carolina State University
c1l98@ncsu.edu

2Johns Hopkins University
c1l98@ncsu.edu

Smoothing of noisy sample covariances is an important component in functional data analysis. We propose a novel covariance smoothing method based on penalized splines and associated software. The proposed method is a bivariate spline smoother that is designed for covariance smoothing and can be used for sparse functional or longitudinal data. We propose a fast algorithm for covariance smoothing using leave-one-subject-out cross validation. Our simulations show that the proposed method compares favorably against several commonly used methods. The method is applied to a study of child growth led by one of coauthors and to a public dataset of longitudinal CD4 counts.

**Modeling Microbial Abundance by Zero-Inflated Semiparametric Gamma Mixed-Effect Models**

*Hongmei Yang, David Topham Xing Qiu, Mary Caserta and Gloria Pryhuber*

**URMC**

Hongmei_Yang@urmc.rochester.edu

The human microbiota are associated with many diseases, such as obesity and diabetes. The advent of high throughput sequencing technology brings convenience to identification of human-associated microbial communities and investigation of their impact on human health, and dynamic microbial data are often collected to study how microbial compositions change over time and how the changes are associated with clinical outcomes. In reality, microbial data often have large proportion of zeros, and are typically highly skewed. Next, the dynamic microbial data are correlated due to the repeated collection of microbial samples from same subjects, and it is important to acknowledge the correlation to yield correct inference. Furthermore, the dynamic profiles of microbiome do not follow a particular form and thus parametric models are inappropriate to describe the profiles. Recently a two-part model named zero-inflated Beta regression (ZIBR) was used to analyze microbial data, with the logistic regression component modelling the inflation of zero and the Beta regression component modelling non-zero relative abundance, and the two parts including the potential clinical predictors independently. A problem with this model is it does not model the longitudinal profiles but consider the correlations among microbiome data from same subjects. We propose the Zero-Inflated Semiparametric Gamma regression with random effects (ZISGR) to address following issues: 1). Model zero inflation by logistic mixed effect models; 2). Model microbial abundance by semiparametric Gamma mixed effect models, to relax parametric assumption about temporal dynamic microbial data, and also to take account of within-subject correlations. The proposed model is applied to the human vaginal microbiota sequence data.

### Session 91: Student session (II)

#### AN ADAPTABLE GENERALIZATION OF HOETLING’S T2 TEST IN HIGH DIMENSION

*HAORAN LI1, Alexander Aue1, Debashis Paul1, Jie Peng1 and Pei Wang2*

1University of California Davis
2School of Medicine at Mount Sinai

hrli@ucdavis.edu

A two-sample test is proposed for detecting differences of mean vectors in a high-dimensional regime based on a ridge-regularized Hotelling’s T2. A method for choosing the regularization parameter that aims to maximize the power within a class of local alternatives is derived. A composite test that combines such optimal tests corresponding to a special collection of local alternatives is also proposed. The cut-off values are derived through an asymptotic analysis of the stochastic process corresponding to the ridge-regularized Hotelling’s T2. Although this test is derived under Gaussianity, the same large sample properties are established under a class of non-Gaussian distributions. Through an extensive simulation study, the composite test is shown to compare favorably against a host of existing two-sample test procedures in a wide range of settings. A different generalization, with a more flexible structure for the test statistic compared to the ridge-regularized Hotelling’s T2 is also proposed.
Adaptive Inferential Method for Monotone Graph Invariants

†Junwei Lu1, Matey Nykov3 and Han Liu2
1Princeton University
2Penn State University, 3Iowa State University
junweil@princeton.edu

We consider the problem of undirected graphical model inference. In many applications, instead of perfectly recovering the unknown graph structure, a more realistic goal is to infer some graph invariants (e.g., the maximum degree, the number of connected subgraphs, the number of isolated nodes). In this paper, we propose a new inferential framework for testing nested multiple hypotheses and constructing confidence intervals of the unknown graph invariants under undirected graphical models. Compared to perfect graph recovery, our methods require significantly weaker conditions. This paper makes two major contributions: (i) Methodologically, for testing nested multiple hypotheses, we propose a skip-down algorithm on the whole family of monotone graph invariants (The invariants which are non-decreasing under addition of edges). We also show that the same skip-down algorithm also provides valid confidence intervals for the targeted graph invariants. (ii) Theoretically, we prove that the length of the obtained confidence intervals are optimal and adaptive to the unknown signal strength. We also prove generic lower bounds for the confidence interval length for various invariants. Numerical results on both synthetic simulations and a brain imaging dataset are provided to illustrate the usefulness of the proposed method.

Inter-Subject Analysis: A Formal Theory

†Cong Ma, Junwei Lu and Han Liu
Princeton University
congm@princeton.edu

This paper develops a formal theory of inter-subject analysis (ISA). ISA aims at exploring the dependency structure between different subjects with the intra-subject dependency as nuisance. Though it has important applications in cognitive neuroscience to locate the stimulus-driven activities, ISA is in lack of a model-based interpretation and is deficient in formal statistical inference. To bridge this gap, a major contribution of this paper is to develop the first modeling framework of ISA based on the Gaussian graphical model. Under this framework, ISA can be converted into the estimation and inference of the inter-subject precision matrix. For estimation, we introduce a novel estimator for the inter-subject dependency whose consistency can be achieved even if the intra-subject dependency is dense. For inference, we propose an untangle and chord procedure to de-bias our estimator. It is valid without the sparsity assumption on the inverse Hessian of the log-likelihood function. This inferential method is general and can be applied to many other statistical problems. Hence it is of independent theoretical interest. Numerical experiments on simulated and brain imaging data validate our methods and theory.

High-Dimensional Function-on-Scalar Regression

†Matthew Reimherr1 and Alice Parodi2
1Penn State University, 2Politecnico di Milano
mreimherr@psu.edu

In this presentation I will discuss recent work concerning function-on-scalar regression when the number of predictors is much larger than the sample size. In particular, I will present a new methodology, called FLAME for Functional Linear Adaptive Mixed Estimation, which simultaneously selects, estimates, and smooths the important predictors in the model. Our methodology is readily available as an R package that utilizes a coordinate descent algorithm for fast implementation. Asymptotic theory will be provided and we will compare to previous methods via simulations and a data application.

Image-on-Scalar Regression via Bivariate Penalized Splines

†Guannan Wang1 and Lily Wang2
1College of William & Mary
2Iowa State University
gwang01@wm.edu

Motivated by recent work of analyzing imaging data in the neuroimaging studies, we consider a class of linear functional response regression models for imaging responses and scalar predictors. To help define the possibly irregular domain of the active (non-null) region, our smoothing method is based on penalized bivariate splines over triangulations, which solves the problem of “leakage” across non-rectangular domains where many conventional smoothing tools suffer. Highly efficient and scalable estimation algorithm is developed. We conduct Monte Carlo simulation to examine the finite-sample performance of the proposed method.

A Multi-Dimensional Functional Principal Components Analysis of EEG Data

Kyle Hasenstab, Aaron Scheffler, Donataello Telesca, Catherine Sugar, Shafali Jeste, Charlotte DiStefano and †Damla Senturk
UCLA
dsenturk@ucla.edu

The electroencephalography (EEG) data created in event-related potential (ERP) experiments have a complex high-dimensional structure. Each ERP trial or, more precisely, switched ERP waveforms which is an instance of functional data. The experiments are made up of sequences of multiple trials, resulting in longitudinal functional data and moreover, responses are recorded at multiple electrodes on the scalp, adding an electrode dimension. Traditional EEG analysis involves multiple simplifications of this structure to increase the signal-to-noise ratio, effectively collapsing the functional and longitudinal components by identifying key features of the ERPs and averaging them across trials. Motivated by an implicit learning paradigm used in autism research in which the functional, longitudinal and electrode components all have critical interpretations, we propose a multidimensional functional principal component analysis (MD-FPCA) technique which does not collapse any of the dimensions of the ERP data. The proposed decomposition is based on separation of the total variation into subject and subunit level variation which are further decomposed in a two-stage functional principal components analysis. The proposed methodology is shown to be useful for modeling longitudinal trends in the ERP functions, leading to novel insights into the learning patterns of children with Autism Spectrum Disorder (ASD) and their typically developing peers as well as comparisons between the two groups. Finite sample properties of MD-FPCA are further studied via extensive simulations.
Utilizing Patient-Level Characteristics for Identification of Cancer Driver Genes

Ho-Hsiang Wu¹, Nilanjan Chatterjee² and Bin Zhu¹
¹National Cancer Institute
²Johns Hopkins University
cossincos@gmail.com

In cancer genomic studies, identifying driver genes associated with cancer initiation and progression is crucial for diagnosis and treatment. For this aim, a probability model is established and a modified one-sided score test is proposed to discover the driver gene with higher frequency of non-silent mutations than its background rate. Distinct from other existing methods assuming homogeneous selection pressures across subjects, the proposed method considers subject-level risk factors and is able to detect selection pressure heterogeneity among subjects. Applied to the lung squamous cell carcinoma of The Cancer Genome Atlas (TCGA) dataset, the proposed method is shown to properly control the false positive rate, identify known cancer driver genes, and discover new ones characterized by the subject-level risk factors.??

Reproducible RNA-seq analysis with recount²

Leonardo Collado Torres¹, Abhinav Nellore², Kai, Shannon Kammers, Ellis³, Margaret Kasper Taub, Hansen³, Andrew Jaffe³, Ben Langmead⁴ and Jeff Leek⁵
¹Lieber Institute for Brain Development
²Comp. Bio. Program and Dept. of Surgery, OHSU
³Department of Biostatistics, JHSPH
⁴Department of Computer Science, JHU
leo.collado@libd.org

In the past decade high-throughput RNA sequencing (RNA-seq) has become the dominant method for studying gene expression. Commonly, researchers generate RNA-seq data to perform a differential expression analysis by either reconstructing transcripts or counting reads that overlap known gene structures. Previously we introduced an intermediate statistical approach called differential expressed region (DER) finder that identifies contiguous regions of the genome showing differential expression signal at single base resolution in an annotation-agnostic manner (Collado-Torres et al, 2017, NAR). These three different approaches all require aligning the RNA-seq data against the reference genome. While over 2,000 projects are publicly available via the Sequence Read Archive (SRA) it is computationally intensive to re-analyze the RNA-seq data in a homogenous way. Using Rail-RNA (Nellore et al, 2016, Bioinformatics) we aligned over 70,000 human RNA-seq samples and created summaries at the gene, exon, exon-exon junction and base pair coverage levels which are available via the recount² resource at https://jhubiostatistics.shinyapps.io/recount² and the recount Bioconductor package (Collado-Torres, Nellore et al, in press). The data from the recount² resource can be used for different levels of RNA-seq differential expression analysis while bypassing the costly computational step of alignment. The data from recount² can be used to predict phenotype information that is otherwise impractical to obtain for all SRA human projects. In recount² we also provide data for the Genotype-Expression Tissue project (GTEx) and The Cancer Genome Atlas (TCGA) which have richer metadata and can be used for developing new methods and testing them before applying them to the SRA data. We expect that recount² will be very useful for the genomics community and in particular for the statistical community by fostering the development of new methods, just like ReCount (Frazee et al, 2011, Bioinformatics) played a roll in the development of DESeq², voom and metagenomeSeq among other methods.

TSNet: a new method for constructing tumor specific gene co-expression networks

Francesca Petralia, Li Wang and Pei Wang
Icahn School of Medicine at Mt Sinai, New York
francesca.petralia@mssm.edu

Tumor tissue samples often contain an unknown fraction of normal cells. This problem well known as tumor purity heterogeneity (TPH) was recently recognized as a severe issue in omics studies. Specifically, if TPH is ignored when inferring co-expression networks, edges are likely to be estimated among genes with mean shift between normal and tumor cells rather than among gene pairs interacting with each other in tumor cells. To address this issue, we propose TSNet a new method which constructs tumor-cell specific gene/protein co-expression networks based on gene/protein expression profiles of tumor tissues. TSNet treats the observed expression profile as a mixture of expressions from different cell types and explicitly models tumor purity percentage in each tumor sample. The advantage of TSNet over existing methods ignoring TPH is illustrated through extensive simulation examples. We then apply TSNet to estimate tumor specific co-expression networks based on breast cancer expression profiles. We identify novel co-expression modules and hub structure specific to tumor cells.

Methods and Applications of Percentile Estimation

Qi Xia¹, Yi Tsong² and Yu-Ting Weng³
¹Temple University
²FDA
tuf14573@temple.edu

The equivalence test in analytical similarity assessment uses a margin of 1.5 times of the standard deviation of reference product. In the current practice, the standard deviation, although estimated using study data, is considered as a fixed constant in the margin. The impacts of such practice, such as the inflation of type I error rate and the reduction of power, were studied before. In 2015 and 2016, various approaches were proposed to address these issues. These proposed approaches include Wald test using restricted maximum likelihood estimate (RMLE) of the standard error, generalized pivotal quantity (GPQ), and exact test based methods. Through simulation comparisons, it is demonstrated that all approaches except Wald test with RMLE increase the type I inflation especially when the sample size is equal and small. We propose a further modification of Wald test with RMLE of the standard error used in our approach. On the other hand, Wald test with RMLE of the standard error could have much low type I error rate when the sample size is equal and small. We propose a further modification of Wald test with RMLE of the standard error in order to improve the type I error rate and power. The critical points for hypothesis testing are simulated for the small sample sizes in order to improve the type I error rate and power from the Wald test approach and Exact based test. The results and discussion through simulation studies will be presented.

Abstracts
Percentile is ubiquitous in statistics and plays a significant role in the day-to-day statistical application. Not only it can be applied to screening and confirmatory cut-point determination in immunogenicity assays, but also the general percentile formulation enriches the statistical literature for mean comparison between reference and test groups in bioequivalence or biosimilarity studies, with the analytical biosimilarity evaluation and scaled average bioequivalence as special cases. Shen et al. (2015) proposed and compared exact based approach with some approximated approaches in one sample scenario for cut-point determination. However, the exact-based approach has the issue of computational time complexity. In this article, we explored more approximated approaches for percentile estimation such as Method of Variance Estimates Recovery (MOVER) based approaches and Modified Large Sample (MLS) approaches. All these approximation approaches are compared with the exact-based approach in one or two samples scenarios. The applications and performance comparison for each approach are displayed with numerical results.

**Statistical Considerations in Demonstrating CMC Analytical Similarity for a Biosimilar Product**

*Richard Burdick¹, Neal Thomas² and Aili Cheng³*

¹Elion Labs, Arizona State University
²Pfizer, Pharmaceutical Sciences and Manufacturing
³Pfizer, Statistical Research and Consulting Center

[Tianhua Wang@fda.hhs.gov](mailto:Tianhua.Wang@fda.hhs.gov)

Demonstration of analytical similarity between a reference product and a biosimilar product is required as part of the biosimilarity approval process. A statistical two one-sided test (TOST) based on the difference of means proposed in the literature and recommended by the FDA for Tier 1 quality attributes is demonstrated to have a type I error rate greater than the specified level, which cannot be corrected by increasing sample size. In this presentation, an alternative TOST based on the effect size is demonstrated to maintain the desired type I error rate, and in some situations, provides greater power than the TOST based on the mean difference. Results are demonstrated both analytically and with computer simulations. An example with calculations is provided.

**SAMPLE SIZE REQUIREMENT FOR ANALYTICAL BIOSIMILARITY ASSESSMENT**

*Tianhua Wang, Yi Tsong and Meiyu Shen*

FDA

[Tianhua.Wang@fda.hhs.gov](mailto:Tianhua.Wang@fda.hhs.gov)

FDA recommends a stepwise approach for obtaining the totality-of-the-evidence for assessing biosimilarity between a proposed biosimilar product and its corresponding reference biologic product be considered (FDA, 2015). The stepwise approach starts with analytical studies for assessing similarity in critical quality attributes which are relevant to clinical outcomes. For critical quality attributes (CQA) that are most relevant to clinical outcomes, FDA requires equivalence test be performed for similarity assessment based on an equivalence acceptance. In practice, the number of CQAs is limited between two and four. The sample size often recommended to be no less than 10 lots per product and needs to be balanced within 150%. Accordingly, we proposed sample size calculation method based on FDA's practical recommendations for the equivalence test currently used in analytical biosimilar assessment.

**Session 95: Statistical Methods for Modelling Data Complexity in Genomics Studies**

**Fast Bayesian Variable Screenings for Binary Response Regressions with Small Sample Size**

*Sheng-Mao Chang¹, Jung-Ying Tzeng² and Ray-Bing Chen¹*

¹National Cheng Kung University
²North Carolina State University

[smchang@mail.ncku.edu.tw](mailto:smchang@mail.ncku.edu.tw)

Screening procedures play an important role in data analysis, especially in high-throughput biological studies where the datasets consist of more covariates than independent subjects. In this article, a Bayesian screening procedure is introduced for the binary response models with logit and probit links. In contrast to many screening rules based on marginal information involving one or a few covariates, the proposed Bayesian procedure simultaneously models all covariates and uses closed form screening statistics. Specifically, we use the posterior means of the regression coefficients as screening statistics; by imposing a generalized g-prior on the regression coefficients, we derive the analytical form of their posterior means and compute the screening statistics without Markov chain Monte Carlo implementation. We evaluate the utility of the proposed Bayesian screening method using simulations and real data analysis. When the sample size is small, the simulation results suggest improved performance with comparable computational cost.

**Multivariate selection models for labelling-based proteomics data with non-ignorable missingness**

*Lin Chen¹, Jiebiao Wang², Pei Wang ² and Don Hedeker¹*

¹University of Chicago
²Mt Sinai Hospital

[lchen@healthbsd.uchicago.edu](mailto:lchen@healthbsd.uchicago.edu)

Mass tag labelling techniques, such as isobaric tags for relative and absolute quantitation (iTRAQ) and tandem mass tags (TMT) have been widely adopted in mass spectrometry experiments. These techniques allow proteins or peptides from multiple samples of a batch to be quantified in a single experiment, but they come at a cost of severe batch effects and non-ignorable missing data occurring at the batch level. In order to jointly analyzing multiple proteins from a priori defined biological pathways, we propose multivariate mixed-effects selection models (MvMSM) to account for the clustered data structure from batch-processed data, and model the unique batch-level missing-data mechanism. Furthermore, we introduce a lasso penalty on the error precision matrix for high-dimensional sparse biological dependence among multiple proteins in a pathway. We employed an alternating direction method of multipliers (ADMM) and an expectation-maximization (EM) algorithm for model estimation. The proposed model framework can also be applied to jointly analyzing multiple phosphopeptides from each phosphoprotein; and we propose a factor-analytic random effects structure for modelling experimental correlations among the phosphopeptides. Simulations demonstrate the advantages of the proposed method in reducing estimation bias, controlling type I error rates and improving power in testing, compared to conventional methods. We apply the proposed MvMSMs to the iTRAQ-based breast cancer proteomics data from the Clinical Proteomic Tumor Analysis Consortium, and identify KEGG pathways and phosphoproteins showing differential abundance levels in triple negative breast cancer (TNBC) subtype versus other subtypes. The proposed method and algorithm can be applied to general multivariate analyses based on clustered data with outcome-dependent missingness.
Joint Modeling and Analysis of Microbiome with Other Omics Data
Xiang Zhan¹, Anna Plantinga², Ni Zhao³ and Michael Wu¹
¹Fred Hutchinson Cancer Research Center  
²University of Washington  
³The Johns Hopkins University
mcwu@fredhutch.org
Understanding the relationship between microbiome and other omics data types is important both for obtaining a more comprehensive view of biological systems as well as for elucidating mechanisms underlying outcomes and response to exposures. However, such analyses are challenging. Issues inherent to microbiome data include dimensionality, compositionality, sparsity, phylogenetic constraints, and complexity of relationships among taxa. It remains unclear how to address these issues, much less to address these issues in combination with problems specific to other omics data types and problems in modeling relationships between microbial taxa and other omics features. To move towards joint analysis, we propose development of methods for studying community level correlations between microbiome and other data types. Real data analyses demonstrate that our approach for correlating microbial taxa with other omics features can reveal new biological findings.

A Non-parametric Framework for High Dimensional Multi-Modal Data With Applications to Imaging Genetics
Yisheng Li¹, Dustin Pluta², Tong Shen¹ and Hernando Ombao²
¹University of California, Irvine  
²University of California, Irvine; KAUST
zhaoxia@ics.uci.edu
Advances in neuroimaging and biotechnology make it possible to investigate genetic perspectives of brain activities in a systematic way. Known as imaging genetics, this area has gained increased attention in multiple scientific disciplines. However, there have been limited studies on statistical methods and computational tools for genetic analysis of functional imaging data. The key challenge here is to model associations between sets of massive, complex and high dimensional data (i.e., genetic and neuroimaging). Here we propose a flexible and easy-to-implement non-parametric framework to examine the effects of genetic variants on brain connectivity captured by functional Magnetic Resonance Imaging (fMRI). The principles that we propose are applicable to various types of high dimensional data. We will also illustrate how the methods are connected to classical regression-based methods.

Session 96: Recent advances in early phase oncology trials using drug combinations
Bayesian dose-finding designs for drug combinations based on time to response and toxicity outcomes
Xiao Su¹ and Yisheng Li²
¹The University of Texas School of Public Health  
²The University of Texas MD Anderson Cancer Center
yeli@mdanderson.org
Most existing dose-finding designs for drug combinations are based on short-term binary or ordinal toxicity and/or efficacy outcomes. We propose a Bayesian phase I/II design based on the time to response and toxicity using an accelerated failure time (AFT) model for the distribution of each time-to-event outcome at each dose combination and the Farlie-Gumbel-Morgenstern (FGM) copula model to account for the dependency between the two outcomes. Rather than assuming a parametric dose-toxicity and dose-response relationship, we only impose partial ordering assumptions on the acceleration factors across dose combinations. Posterior inference is made using Bayesian isotonic regression transformation of unordered posterior samples of the model parameters. A utility function is constructed based on trade-offs between the two time-to-event outcomes as a result of consultation with the physicians. We propose a dose-finding algorithm to adaptively allocate new patients to a safe and efficacious combination with the probability proportional to the posterior median utility associated with that combination. At the end of the trial, an optimal combination is selected that maximizes the posterior median utility while being considered safe and not futile. Our simulation study suggests that under a variety of scenarios, the proposed method has a high probability of selecting the optimal dose combination and treating most patients at desirable dose combinations.

A curve-free Bayesian decision-theoretic design for two-agent phase I trial and its extension
Ying Lu¹, Bee Leng Lee² and Hua Jin³
¹Stanford University  
²San Jose State University  
³South China Normal University
yliu@stanford.edu
In this paper, we present a curve-free Bayesian decision model to select MTD for phase I trials for combination of two drugs under a partially monotonic toxicity assumption. We will then discuss on its extension to evaluate both safety and biological activities in a phase I trial for targeted agents and immunotherapy under a constrain that the probability of biomarker activities is a unimode function of dose. We demonstrate our method through simulations.

Session 97: Advanced Statistical Methods for Complicated High Dimensional Data
A new approach for identifying and removing the cell-cycle effect from single-cell RNA-seq data
Martin Barron and Jun Li
University of Notre Dame
jun.li@nd.edu
Single-cell RNA-Sequencing (scRNA-Seq) is a revolutionary technique for discovering and describing cell types in heterogeneous tissues, yet its measurement of expression often suffers from large systematic bias. A major source of this bias is the cell cycle, which introduces large within-cell-type heterogeneity that can obscure the differences in expression between cell types. The current method for removing the cell-cycle effect is unable to effectively identify this effect and has a high risk of removing other biological components of interest, compromising downstream analysis. We present ccRemover, a new method that reliably identifies the cell-cycle effect and removes it. ccRemover preserves other biological signals of interest in the data and thus can serve as an important pre-processing step for many scRNA-Seq data analyses. The effectiveness of ccRemover is demonstrated using simulation data and three real scRNA-Seq datasets, where it boosts the performance of existing clustering algorithms in distinguishing between cell types.

Robust Normalization of High-throughput Proteomics Experiments
Hua Tang, Meng Wang, Huaying Fang, Lihua Jiang, Shin Lin and Michael Snyder
Stanford University
huatang@stanford.edu
The University of Texas MD Anderson Cancer Center
Shotgun proteomics, based on liquid chromatography coupled with tandem mass spectrometry (LC-MS), have enabled high-throughput and quantitative characterization of protein variation in biological samples. However, normalization of such data poses statistical challenges. We developed a normalization approach, which accounts for technical variation between mass-spectrometry runs and between biological samples. By adaptively identifying a subset of proteins, whose abundances are relativelystable across samples, our model allows other proteins to vary substantially. This feature is particularly important when analyzing biological samples from differentorganisms or different tissues. We evaluate this method on simulated and real data.

On Consistency of Graph-based Semi-supervised Learning
Chengan Du and Yunpeng Zhao
George Mason University

Graph-based semi-supervised learning is one of the most popular methods in machine learning. Some of its theoretical properties such as bounds for the generalization error and the convergence of the graph Laplacian regularizer have been studied in computer science and statistics literatures. However, a fundamental statistical property, the consistency of the estimator from this method has not been proved. In this talk, we study the consistency problem under a non-parametric framework. We prove the consistency of graph-based learning in the case that the estimated scores are enforced to be equal to the observed responses for the labeled data. The sample sizes of both labeled and unlabeled data are allowed to grow in this result. When the estimated scores are not required to be equal to the observed responses, a tuning parameter is used to balance the loss function and the graph Laplacian regularizer. We give a counterexample demonstrating that the estimator for this case can be inconsistent. The theoretical findings are supported by numerical studies.

QRank: A novel quantile regression tool for eQTL discovery
Xiaoyu Song, Gen Li, Zhenwei Zhou, Xianling Wang, Iuliana Ionita-Laza and Ying Wei
Columbia University

Motivation: Over the past decade, there has been a remarkable improvement in our understanding of the role of genetic variation in complex human diseases, especially via genome-wide association studies. However, the underlying molecular mechanisms are still poorly characterized, impeding the development of therapeutic interventions. Identifying genetic variants that influence the expression level of a gene, i.e. expression quantitative trait loci (eQTLs), can help us understand how genetic variants influence traits at the molecular level. While most eQTL studies focus on identifying mean effects on gene expression using linear regression, evidence suggests that genetic variation can impact the entire distribution of the expression level. Motivated by the potential higher order associations, several studies investigated variance eQTLs.

Results: In this paper, we develop a Quantile Rank-score based test (QRank), which provides an easy way to identify eQTLs that are associated with the conditional quantile functions of gene expression. We have applied the proposed QRank to the Genotype-Tissue Expression project, an international tissue bank for studying the relationship between genetic variation and gene expression in human tissues, and found that the proposed QRank complements the existing methods, and identifies new eQTLs with heterogeneous effects across different quantile levels. Notably, we show that the eQTLs identified by QRank but missed by linear regression are associated with greater enrichment in genome-wide significant SNPs from the GWAS catalog, and are also more likely to be tissue specific than eQTLs identified by linear regression.

Availability and Implementation: An R package is available on R CRAN at https://cran.r-project.org/web/packages/QRank.

Session 98: Advances in Low-rank Modeling and Its Estimation

Recent Advances in Collaborative Ranking
Cho-Jui Hsieh
UC Davis

Matrix factorization algorithms have become one of the most popular approaches for building large-scale recommender systems. However, in many applications the goal is to generate personalized rankings instead of predicting the absolute ratings. For example, in movie recommendation systems, the goal is to predict top-k movies for each user, instead of predicting all the unobserved user-movie ratings. For these problems, we discuss an effective approach, called collaborative ranking, and show our recent work on solving large-scale collaborative ranking problems, and apply it to real datasets.

Nonparametric Operator-Regularized Covariance Function Estimation
Raymond K. W. Wong1 and Xiaoke Zhang2
1Iowa State University
2University of Delaware

This paper develops a class of nonparametric covariance function estimators by utilizing spectral regularization of an operator, which is associated with a typically infinite dimensional reproducing kernel Hilbert space. By construction, these estimators are positive semi-definite and hence valid covariance functions. A related representer theorem is established to provide a finite dimensional representation of such estimators. In order to achieve low-rank estimations, trace-norm regularization is studied in detail. A specific computational algorithm is developed and this estimator is shown to enjoy excellent rates of convergence under either fixed or random designs. The empirical performance of the proposed trace-norm-regularized estimator is demonstrated in a simulation study, while its practical utility is illustrated in an analysis of a traffic data set. This is joint work with Xiaoke Zhang.

Parsimonious matrix-variate regressions through envelope models
Shanshan Ding1 and Dennis Cook2
1University of Delaware
2University of Minnesota

Envelopes (Cook et al., 2010) were originally proposed for multivariate regression analysis to achieve potential efficiency gains via dimension reduction. In this talk, we introduce envelope models for efficient estimation in matrix-variate analysis, where the response Y is a random matrix and the predictor X can be either a scalar, or a vector, or a matrix, treated as non-stochastic. The matrix-variate regression models and their envelope structures are generalizations of conventional regression methods and their envelopes. They can be applied to neuroimaging, cross-over design analysis, temporal and spatial data, and multivariate growth curve modeling, among others.
Under the envelope framework, redundant information can be eliminated in estimation and the number of parameters can be notably reduced when the matrix-variate dimension is large. Therefore, the estimation can be much more accurate. We further investigate the asymptotic properties of the envelope estimators and show their efficiency in both theoretical and numerical studies.

MM Algorithms For Variance Components Models

Hua Zhou1, Liuyi Hu2, Jin Zhou2 and Kenneth Lange3
1 UCLA
2 NCSU
3 University of Arizona
Huazhou@ucla.edu

Variance components estimation and mixed model analysis are central themes in statistics with applications in numerous scientific disciplines. Despite the best efforts of generations of statisticians and numerical analysts, maximum likelihood estimation and restricted maximum likelihood estimation of variance component models remain numerically challenging. In this talk, we present a novel iterative algorithm for variance components estimation based on the minorization-maximization (MM) principle. MM algorithm is trivial to implement and competitive on large data problems. The algorithm readily extends to more complicated problems such as linear mixed models, multivariate response models possibly with missing data, maximum a posteriori estimation, and penalized estimation. We demonstrate, both numerically and theoretically, that it converges faster than the classical EM algorithm when the number of variance components is greater than two.

Session 99: Recent Advancement and Applications in Multiplicity Adjustments

Large-Scale Heterogeneity Testing under Sparsity

Xiang Lyu and Guang Cheng
Department of Statistics, Purdue University
alanlvru@gmail.com

Massive data often consists of a growing number of potentially heterogeneous sub-populations. In this paper, we propose a set of testing procedures for detecting heterogeneity in the framework of high dimensional linear models. These bootstrap-assisted testing procedures can be efficiently implemented in a parallel fashion. In theory, we prove that our large-scale heterogeneity testing is asymptotically consistent and minimax optimal in the sense that it can consistently detect departure from null of a magnitude that no other tests could improve. These theoretical results hold, in particular, when model dimensionality grows exponentially fast and the number of sub-population diverges, and apply to non-Gaussian model errors. Our results are further extended to accommodate heteroscedastic errors commonly seen in massive datasets. Simulations are provided to back up our theoretical results.

A Unified Framework For Weighted Parametric Multiple Test Procedures

Dong Xi1, Ekkehard Glimm2, Willi Maurer2 and Frank Bretz2
1 Novartis Pharmaceuticals Corporation
2 Novartis AG
dong.xi@novartis.com

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 alpha. This may be also necessary to take certain study situations into account. For such cases we introduce a subclass of exact alpha-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.

Robustness of Multiple Testing Procedures in Confirmatory Clinical Trials

Michael Rosenblum, Chenguang Wang and Yuchen Yang
Johns Hopkins University
yangy1608@gmail.com

Title: Robustness of multiple testing procedures in confirmatory clinical trials that test for treatment effects in multiple populations

Abstract: In planning a clinical trial, assumptions are often made in order to determine the sample size and study duration. For example, one may make assumptions about the outcome variance, enrollment rate, etc. We focus on confirmatory trials involving hypothesis tests for two predefined subpopulations and the overall population. These designs require even more a priori assumptions, such as the relative sizes of the subpopulations. We investigate the relative impact of assumption violations on different multiple testing procedures. We explore the tradeoffs in using optimized multiple testing procedures versus simpler procedures, in terms of required sample size and robustness to assumption violations. Specifically, we compare the following procedures: Bonferroni using equal alpha allocation to each null hypothesis, an optimized Bonferroni procedure that allows unequal alpha allocation, and a graphical approach of Maurer and Bretz (2009) that involves reallocation from rejected null hypotheses to the remaining null hypotheses. We demonstrate the efficiency versus robustness trade-off comparing these procedures.

Session 100: Design and analysis of cancer immunotherapy trials

Practical considerations in designing immunotherapy trials with delayed treatment effects

Alan Chiang
Eli Lilly and Company
chiangay@lilly.com

Achieving Optimal Power of Log-rank Test with Random Treatment Time-lag Effect

Zhenzhen Xu1, Yongsook Park2, Boguang Zhen3 and Bin Zhu4
1 FDA/CBER
2 University of Pittsburgh
3 National Cancer Institute
zhengxq@fda.hhs.gov

In some clinical settings such as the cancer immunotherapy trials, a treatment time-lag effect is present and the lag duration may vary from subject to subject or from study to study. An efficient study design would not only take into account the time-lag effect but also consider the individual heterogeneity in the lag duration. In this paper, we present a Generalized Piecewise Weighted Logrank test, designed to account for the random time-lag effect while maximizing the study power in terms of the weight allocation. Based on
Flexible Survival Model and Sample Size Determination in Cancer Vaccine Studies
Takahiro Hasegawa
Shionogi & Co., Ltd.

In recent years, immunological science has evolved, and cancer vaccines are now approved and available for treating existing cancers. Cancer vaccines require time to elicit an immune response and a delayed treatment effect is expected. Because of this delayed effect, the drug does not initially influence the survival curves of the trial results. If the treatment is effective, the survival curves will separate once the vaccine’s effect has established. The delayed treatment effect violates the assumptions of the Cox proportional hazards model and substantially reduces the statistical power of conventional methods such as the standard log-rank test. Therefore, the use of a weighted log-rank test with the Fleming-Harrington class of weights has been proposed as an alternative analysis method for survival endpoints. First of all, we give a flexible survival model in cancer vaccine studies that considers not only the delayed treatment effect but also the long-term survivors which were shown in some cancer vaccine studies. Moreover, we present a method for calculating the sample size for the weighted log-rank test with the Fleming-Harrington class of weights under assumption of the proposed flexible survival model. The impact of delayed effect timing on both the choice of the Fleming-Harrington’s weights and the increment in the required number of events is discussed.

Session 101: Design and practical considerations for dose selection trials

Dose Finding Based on Integrated and Adaptive Approaches
Lei Nie and Rajeshwari Sridhara
The US FDA

The primary role of Phase I clinical trials is to assess the relationship between dose and toxicity and/or the relationship between dose and pharmacokinetic parameters and recommend doses for phase II studies. In oncology, Phase I studies are usually designed to determine the MTD of cytotoxic on the basis of dose limiting toxicity. In the era of targeted therapies including immunotherapy, the traditional approaches are no longer appropriate. Innovative approaches are needed. In this presentation, we discuss the integrated approach through Bayesian design and seamless adaptive design.

Practical Experiences with Phase 1 Dose Escalation Studies
Inna Perevozskaya and Yuehui Wu

Last decade has seen an increased use of innovative adaptive designs in phase 1 oncology trials. These methods were originally developed as an alternative to the traditional 3+3 design, all aiming to improve the precision of Maximum Tolerated Dose (MTD) finding. Among the multitude of “alternative” methods available today, Toxicity Probability Interval (TPI) and Continued Reassessment Method (CRM) stand out. Both are commonly used Bayesian adaptive designs to estimate the MTD in Phase 1 oncology clinical trials. Modified versions of both methods (mTPI and mCRM) are considered here. Patients are enrolled in cohorts of variable size (2 to 4), starting with the lowest dose for the first cohort. Subsequent dose escalations are limited to a maximum 100% increase from the previously studied dose. Both algorithms utilize knowledge about dose-toxicity relationship derived from the interim data to assign next cohort of patients to the current estimate of MTD. But they handle the data differently: mTPI models probability of toxicity at each dose separately, while CRM relies on parametric dose-toxicity model pulling information across the doses. The study utilizing either mTPI or mCRM will continue accruing patients until either maximum sample size is reached or MTD is found with sufficient accuracy or all doses appear to be unacceptably toxic. In this talk, we will look at some case-study motivated examples to compare the operating characteristics and implementation experiences with both methods.

Application of Bayesian Methods in Oncology Dose Escalation Studies with Late Onset Toxicity
Li Liu
Sanofi

In phase I oncology trials, dose-limiting toxicity (DLT) is often used as the endpoint in dose escalation studies to find the maximum tolerated dose (MTD). Bayesian adaptive designs with overdose control have been introduced to find the MTD while protecting patients from overdosing. However, these methods may not be feasible when a particular adverse event is likely to occur during long term exposure. Tighiouart, Liu and Rogatko (2014) proposed an Escalation with Overdose Control (EWOC) design using a time to event toxicity endpoint, which can accommodate a longer assessment time while maintaining escalation timelines. In this talk, Tighiouart’s time to event EWOC method has been applied with some modifications based on practical considerations for a dose escalation study in which both short term DLTs and a long term adverse event of special interest are safety concerns. The operating characteristics of the modified time to event EWOC design are explored through simulations, and some practical issues in implementation are addressed.

Longitudinal Data Generation Based on Bootstrapped Samples from an Observational Study with Application
Greg Cicconetti1, Deli Wang2 and Weinong Robieson

To support the design and interim analysis planning in an Alzheimer’s disease clinical trial, extensive simulations are required in order to evaluate the operating characteristics of interim Go/No-Go decision rules. A pivotal first step, particularly when interim analyses will leverage longitudinal patient data, is to simulate data that is realistic. The endpoint of interest, Clinical Dementia Rating Scale-Sum of Boxes (CDR-SB), is measured at baseline, 6, 12, 18, and 24 months and realizes half-integer values from \{0, 0.5, 1.0, \ldots, 18\}. There are limitations to simulate CDR-SB using the normal approximation method due to special values taken for CDR-SB. In order to generate placebo data that reflects longitudinal disease progression, it is common to lean on observational data for assumptions. In this case, data from Alzheimer’s Disease Neuroimaging Initiative serves as our starting point. The workflow first includes imputation of the observational data set so that we have
Session 102: Novel Statistical Methods for Analysis of Complex Biological Data

Accounting for Within-gene Correlation Structure in RNA-seq Differential Expression Analysis
Yet Nguyen and Dan Nettleton
Iowa State University
dnett@iastate.edu

RNA-seq technology has become the primary approach for high-throughput measurement of transcript abundance levels. The technique is now commonly used in experiments with complex designs that yield dependent measures of expression within each gene. Although there are many methods for RNA-seq differential expression analysis, existing methods do not properly account for within-gene correlations. We address this shortcoming by using normalized log-counts and associated precision weights in a general linear model pipeline that relies on parametric bootstrap techniques to conduct valid and powerful differential expression inference. Simulation studies show the advantages of our method over alternatives that do not account for correlation among observations within each gene. Example analyses for split-plot and repeated-measures RNA-seq experiments will be discussed.

Tensor-on-tensor Regression
Eric Lock
University of Minnesota
eLOCK@umn.edu

We propose a framework for the linear prediction of a multi-way array (i.e., a tensor) from another multi-way array of arbitrary dimension, using the contracted tensor product. This framework generalizes several existing approaches, including methods to predict a scalar outcome from a tensor, a matrix from a matrix, or a tensor from a scalar. We describe an approach that exploits the multiway structure of both the predictors and the outcomes by restricting the coefficients to have reduced CP-rank. We propose a general and efficient algorithm for penalized least-squares estimation, which allows for a ridge ($L_2$) penalty on the coefficients. The objective is shown to give the mode of a Bayesian posterior, which motivates a Gibbs sampling algorithm for inference. We illustrate the approach with an application to multi-tissue gene expression data.

Bayesian Semiparametric Mixed Effects Markov Chains
Abhra Sarkar and David Dunson
Duke University
abhra.stat@gmail.com

Studying the neurological, genetic and evolutionary basis of human vocal communication mechanisms using animal vocalization models is an important field of neuroscience. The data sets typically comprise structured sequences of syllables or “songs” produced by animals from different genotypes under different social contexts. We develop a novel Bayesian semiparametric framework for inference in such data sets. Our approach is built on a novel class of mixed effects Markov transition models for the songs that accommodate exogenous influences of genotype and context as well as animal-specific heterogeneity. Crucial advantages of the proposed approach include its ability to provide insights into key scientific queries related to global and local influences of the exogenous predictors on the transition dynamics via automated tests of hypotheses. The methodology is illustrated using simulation experiments and the aforementioned motivating application in neuroscience.

Computation of Ancestry Scores with Mixed Families and Unrelated Individuals
Yi-Hui Zhou, J.S. Marron and Fred Wright
1 North Carolina State University
2 University of North Carolina
yihui_zhou@ncsu.edu

The issue of robustness to family relationships in computing genotype ancestry scores such as eigenvector projections has received increased attention in genetic association, and is particularly challenging when sets of both unrelated individuals and closely related family members are included. The current standard is to compute loadings (left singular vectors) using unrelated individuals and to compute projected scores for remaining family members. However, projected ancestry scores from this approach suffer from shrinkage toward zero. We consider two main novel strategies: (i) matrix substitution based on decomposition of a target family-orthogonalized covariance matrix, and (ii) using family-averaged data to obtain loadings. We illustrate the performance via simulations, including resampling from 1000 Genomes Project data, and analysis of a cystic fibrosis dataset. The matrix substitution approach has similar performance to the current standard, but is simple and uses only a genotype covariance matrix, while the family-average method shows superior performance. Our approaches are accompanied by novel ancillary approaches that provide considerable insight, including individual-specific eigenvalue scree plots.

Session 103: TBD

Empirical Bayes and a flexible multiple-testing framework for large-scale simultaneous inference
Zhi Wei
New Jersey Institute of Technology
zhiwei@njit.edu

We propose an empirical Bayes change-point model to identify alternative splicing patterns from RNAseq data. Compared with previous methods, our approach has several unique merits. First of all, our model does not rely on annotation information. Instead, it provides for the first time a systematic framework to integrate various information when available, in particular the useful junction read information, in order to obtain better performance. Second, we utilize an empirical Bayes model to efficiently pool information across genes to improve detection efficiency. Third, we provide a flexible multiple-testing framework in which the user can choose to address different levels of questions, namely, whether alternative splicing happens, and/or where it happens. Simulation studies and real data application have demonstrated that our method is powerful and accurate.

Flexible Spectral Methods for Community Detection in Networks
Pengsheng Ji
University of Georgia
psji@uga.edu

We propose a class of flexible spectral methods for community detection in directed and undirected networks. These methods extract the clustering information by taking the entry-wise ratios of
Yi Ding, Model
High Dimensional Minimum Variance Portfolio under Factor
0 and h(X) at the endpoints 0 and 1, respectively.

We show that hX(??) is a continuous function of ?? over the range

[0,1], thereby providing a smooth interpolation between the values

determined by the population canonical variates and that determined by

the estimated canonical variates. Both losses capture some essence of

CCA, particularly for the cases where the leading canonical correlations are close to 1. Next, we derive new non-asymptotic prediction upper bounds, where (1 − λk 1) (1 − λk+1) appears in the first order term. Besides, we first separate p1 and p2 for the first order term of the upper bounds without assuming the residual correlations are zeros. Moreover, we also show that when the sample size is large enough, sample CCA enjoys the minimax convergence rates under these two losses for a broad class of parameter spaces, including

the cases where the leading canonical correlations are close to 1. Both upper and lower bounds require novel proof techniques. Given many fast CCA algorithms have been proposed in the recent literature, we believe that our rates results are useful as a benchmark reference to fully understand the computational and statistical trade-off for CCA in the future.

Abstracts

Session 104: Analysis of High-Dimensional Data and Applications

Intrinsic Entropies of Log-concave Distributions
Varun Jog
University of Wisconsin - Madison
vjjog@wisc.edu

The entropy of a random variable is well-known to equal the expo-

nal growth rate of the volumes of its typical sets. In this paper,

we show that for any log-concave random variable X, the sequence

of the ???n?????th intrinsic volumes of the typical sets of X in di-

mensions n ??? 1 grows exponentially with a well-defined rate. We
denote this rate by hX(??), and call it the ??th intrinsic entropy of X.

We show that hX(??) is a continuous function of ?? over the range

[0,1], thereby providing a smooth interpolation between the values

of the !!!n!!!!th intrinsic volumes of the typical sets of X in di-

We study tests for high-dimensional covariance matrices when data

exhibit heteroscedasticity. The observation is modeled as a prod-

uct between a heteroscedastic scalar and an iid multidimensional ran-

dom vector. We aim to test the covariance matrix up to the heteroscedastic scalar. To remove the heteroscedasticity, we self-

ormalize the observations and establish a CLT for the linear spec-

tral statistics (LSS) of the sample covariance matrix based on the self-

ormalized observations. The CLT differs from the existing

CLTs for the LSS of the usual sample covariance matrix (Bai and


on the new CLT neither assume a specific parametric distribution for the data nor involve extra terms containing the fourth moment. Em-

pirically, we use our tests to evaluate different predictors of the co-

variance matrix of the S&P 500 Financials sector stock returns. The

results show that: 1) self-normalizing the observations can improve

the prediction; 2) compared with sparsity, an approximate factor model is more suitable for stock returns.

Canonical Correlation Analysis: New Losses And New Rates
Zhuang Ma1 and Xiaodong Li2
1University of Pennsylvania
2University of California, Davis
xdgli@ucdavis.edu

Canonical Correlation Analysis has been proposed for more than 80

years, but the non-asymptotic minimax convergence rates have not

been well studied. In this paper, we propose two new prediction loss

functions based on the mismatching between the model space deter-

mined by the population canonical variates and that determined by the estimated canonical variates. Both losses capture some essence of CCA, particularly for the cases where the leading canonical correlations are close to 1. Next, we derive new non-asymptotic prediction upper bounds, where (1 − λk 1) (1 − λk+1) appears in the first order term. Besides, we first separate p1 and p2 for the first order term of the upper bounds without assuming the residual correlations are zeros. Moreover, we also show that when the sample size is large enough, sample CCA enjoys the minimax convergence rates under these two losses for a broad class of parameter spaces, including the cases where the leading canonical correlations are close to 1. Both upper and lower bounds require novel proof techniques. Given many fast CCA algorithms have been proposed in the recent literature, we believe that our rates results are useful as a benchmark reference to fully understand the computational and statistical trade-off for CCA in the future.

Canonical Correlation Analysis: New Losses And New Rates
Zhuang Ma1 and Xiaodong Li2
1University of Pennsylvania
2University of California, Davis
xdgli@ucdavis.edu

Canonical Correlation Analysis has been proposed for more than 80

years, but the non-asymptotic minimax convergence rates have not

been well studied. In this paper, we propose two new prediction loss

functions based on the mismatching between the model space deter-

mined by the population canonical variates and that determined by the estimated canonical variates. Both losses capture some essence of CCA, particularly for the cases where the leading canonical correlations are close to 1. Next, we derive new non-asymptotic prediction upper bounds, where (1 − λk 1) (1 − λk+1) appears in the first order term. Besides, we first separate p1 and p2 for the first order term of the upper bounds without assuming the residual correlations are zeros. Moreover, we also show that when the sample size is large enough, sample CCA enjoys the minimax convergence rates under these two losses for a broad class of parameter spaces, including the cases where the leading canonical correlations are close to 1. Both upper and lower bounds require novel proof techniques. Given many fast CCA algorithms have been proposed in the recent literature, we believe that our rates results are useful as a benchmark reference to fully understand the computational and statistical trade-off for CCA in the future.
Session 105: Recent advances in clinical trial design and analysis with time to event endpoints

The Win-ratio Statistic in Clinical Trials with Multiple Types of Event
David Oakes
University of Rochester
oakes@bst.rochester.edu

Pocock (2012), following Finkelstein and Schoenfeld (1999) has popularized the “win ratio” as a simple method of statistical analysis for controlled clinical trials with multiple outcomes. The approach uses pairwise comparisons between patients in the treatment and control groups using a primary outcome (say time to cardiac death) with ties broken using a secondary outcome (say time to a nonfatal cardiac event). In general the observed pairwise preferences and the weight they attach to the component rankings will depend on the distribution of potential follow-up time. We present expressions for the “win” and “loss” probabilities for general bivariate survival models when follow-up is limited to a specific time horizon. In the special case of a bivariate Lehmann model we show that the win ratio does not depend on this horizon. We show how the win ratio may be estimated non-parametrically or from a parametric model. Extensions to events of three or more types are described. A novel application of the marginal method of estimation is discussed. Reference. Oakes (2016) Biometrika, 103, 742-745.

Semi-parametric Density Ratio Modeling of Survival Data From a Prevalent Cohort
Hong Zhu1, Jing Ning2, Yu Shen2 and Jing Qin3
1University of Texas Southwestern Medical Center
2University of Texas MD Anderson Cancer Center
3National Institute of Allergy & Infectious Disease
hong.zhu@utsouthwestern.edu

In this work, we consider methods for assessing covariate effects on survival outcome in the target population when data are collected under prevalent sampling. We investigate a flexible semiparametric density ratio model without the constraints of the constant disease incidence rate and discrete covariates as required in Shen, Ning and Qin (2012). For inference, we introduce two likelihood approaches with distinct computational algorithms. We first develop a full likelihood approach to obtain the most efficient estimators by an iterative algorithm. Under the density ratio model, we exploit the invariance property of uncensored failure times from the prevalent cohort and also propose a computationally convenient estimation procedure that uses a conditional pairwise likelihood. The empirical performance and efficiency of the two approaches are evaluated through simulation studies. The proposed methods are applied to the Surveillance, Epidemiology and End Results (SEER)-Medicare linked data for women diagnosed with stage IV breast cancer.

A Bayesian Nonparametric Model to Predict the Time of Hospital Readmission
Song Zhang
University of Texas Southwestern Medical Center
song.zhang@utsouthwestern.edu

Electronic medical records (EMR) systems are being rapidly implemented throughout U.S. Health care organizations. We extend the Cox survival model with a Bayesian additive regression tree (BART) to take advantages of the rich information continuously collected and made available through the EMR system, including diagnostics, interventions, lab measurements and vital signs. The BART component is flexible enough to account for nonlinear effects and high-order interactions. The setup of the Bayesian survival model also allows researchers to model hierarchical structure in the data, such as clustering by hospital and physicians. A simulation study and a real application example are presented.

Session 106: Methods and applications in clustered and high dimensional data

Bayesian multivariate skew meta-regression models for individual patient data
Joseph Ibrahim1, Sungdak Kim2, Ming-Hui Chen3, Arvind Shah4 and Jianxin Lin4
1University of North Carolina
2National Cancer Institute
3University of Connecticut
4Merck Research Laboratories
ming-hui.chen@uconn.edu

We examine a class of multivariate meta-regression models in the presence of Individual Patient Data (IPD). The methodology is well motivated from several studies of cholesterol lowering drugs where the goal is to jointly analyze the multivariate outcomes, Low Density Lipoprotein Cholesterol (LDL-C), High Density Lipoprotein Cholesterol (HDL-C), and Triglycerides (TG). These three continuous outcome measures are correlated and shed much light on a subject’s lipid status. One of the main goals in lipid research is the joint analysis of these three outcome measures in a meta-regression setting. Since these outcome measures are not typically multivariate normal, one must consider classes of distributions that allow for skewness in one or more of the outcomes. In this paper, we consider a new general class of multivariate skew distributions for multivariate meta-regression and examine their theoretical properties. Using these distributions, we construct a Bayesian model for the meta-data and develop an efficient Markov chain Monte Carlo (MCMC) computational scheme for carrying out the computations. In addition, we develop a multivariate L measure for model comparison, Bayesian residuals for model assessment, and a Bayesian procedure for detecting outlying trials. The proposed multivariate L measure, Bayesian residuals, and Bayesian outlying trial detection procedure are particularly suitable and computationally attractive in the multivariate meta-regression setting. A detailed case study demonstrating the usefulness of the proposed methodology is carried out in an IPD multivariate meta-regression setting using 26 pivotal Merck clinical trials that compare statins (cholesterol lowering drugs) in combination with ezetimibe and statins alone on treatment-naive patients and those continuing on statins at baseline.

Dimension Reduction in High Dimensional Multivariate Time Series Analysis
Zeda Li and William W.S. Wei
Dept. of Statistical Science, Temple University
wwei@temple.edu

The vector autoregressive (VAR) and vector autoregressive moving average (VARMA) models have been widely used to model multivariate time series, because of their capability to represent the dynamic relationships among variables in a system and their usefulness in forecasting unknown future values. However, when the dimension is very high, the number of parameters often exceed the number of available observations, and it is impossible to estimate the parameters. A suitable solution is clearly needed. After introducing some existing methods, we will suggest the use of aggregation as a dimension reduction method, which is very natural and
simple to use. We will compare our proposed method with other existing methods in terms of forecasting aggregates. Forecasts accuracy is evaluated through both simulations and empirical examples.

Semiparametric Monitoring test based on clustered data
Jiahua Chen1, Pengfei Li2 and Yukan Liu3
1Dept of Statistics, University of British Columbia
2Dept of Statistics, University of Waterloo
3School of Mathematics and Statistics, ECNU

Due to factors such as climate change, forest fire, plague of insects on lumber quality, it is important to update (statistical) procedures in American Society for Testing and Materials (ASTM) Standard D1990 (adopted in 1991) from time to time. The statistical component of the problem is to detect the change in the lower percentiles of the solid lumber strength. Eight statistical tests are studied by Verrill et al. (2015) to determine if they perform acceptably when applied to test data from a monitoring program. Some well-known methods such as Wilcoxon and Kolmogorov-Smirnov tests are found to have severely inflated type I errors when the data are clustered. A new method that performs well in the presence of random effects is therefore in urgent need. In this paper, we develop a novel test by combining composite empirical likelihood, cluster-based bootstrapping and density ratio model. The test satisfactorily controls the type I error in monitoring the trend of lower quantiles and conclusions are supported by asymptotic results. Our results are generic, not confined to wood industry applications.

Automatic Region-wise Spatially Varying Coefficient Regression Model
Shuo Chen and Chengsheng Jiang
University of Maryland, College Park

cjiang89@umd.edu

It is well known that particular matter 2.5 (PM 2.5) is associated with adverse health outcome. Recent years, it is found that the different chemicals composition of PM2.5 has different impacts on health. Based on national dataset of PM2.5 and cardiovascular disease mortality rate (CVD) for 3109 counties, we develop a novel statistical framework to automatically identify spatially varied association between PM2.5 and CVD within certain regions. We calculate a regional index for each county which will be used to generate regional boundaries. The regional index shows how the impact of this county on the overall association between PM2.5 and CVD for all counties. Then, we segment the whole map into a set of disjoint regions based on the spatial adjacency matrix with constraints that all spatial units within a region are contiguous and have similar association between exposure and health outcome. We implement the framework by using regression and spectral graph techniques. We develop goodness of fit criteria for model assessment and model selection. The simulation study confirms the satisfactory performance of our model. The further investigation of association between PM2.5 and CVD confirmed spatial varying associations. Especially, it revealed that the highest PM2.5 risk on CVD is neither in high exposure nor in the high mortality of CVD area. This result prove insightful guidance for environmental health research and evidence for government to apply different control policy for PM2.5.

Session 107: Innovative clinical trial designs

A Calibrated Power Prior Approach to Borrow Information from Historical Data with Application to Bio
Haitao Pan1, Ying Yuan1 and Jielai Xia2
1The University of Texas MD Anderson Cancer Center
2Fourth Military Medical University
davidhaitaoapan@gmail.com

A biosimilar refers to a follow-on biologic intended to be approved for marketing based on biosimilarity to an existing patented biological product (i.e., the reference product). To develop a biosimilar product, it is essential to demonstrate biosimilarity between the follow-on biologic and the reference product, typically through two-arm randomization trials. We propose a Bayesian adaptive design for trials to evaluate biosimilar products. To take advantage of the abundant historical data on the efficacy of the reference product that is typically available at the time a biosimilar product is developed, we propose the calibrated power prior, which allows our design to adaptively borrow information from the historical data according to the congruence between the historical data and the new data collected from the current trial. We propose a new measure, the Bayesian biosimilarity index, to measure the similarity between the biosimilar and the reference product. During the trial, we evaluate the Bayesian biosimilarity index in a group sequential fashion based on the accumulating interim data, and stop the trial early once there is enough information to conclude or reject the similarity. Extensive simulation studies show that the proposed design has higher power than traditional designs. We applied the proposed design to a biosimilar trial for treating rheumatoid arthritis.

A Nonparametric Bayesian Basket Trial Design
Yanxun Xu1, Peter Mueller2, Apostolia Tsimberidou3 and Donald Berry3
1Johns Hopkins University
2The University of Texas at Austin
3The University of Texas MD Anderson Cancer Center
yxu.stat@gmail.com

Targeted therapies on the basis of genomic aberrations analysis of the tumor have shown promising results in cancer prognosis and treatment. Regardless of tumor type, trials that match patients to targeted therapies for their particular genomic aberrations have become a mainstream direction of therapeutic management of patients with cancer. Therefore, finding the subpopulation of patients who can most benefit from an aberration-specific targeted therapy across multiple cancer types is important. We propose an adaptive Bayesian clinical trial design for patient allocation and subpopulation identification. We start with a decision theoretic approach, including a utility function and a probability model across all possible subpopulations. The main features of the proposed design and population finding methods are that we allow for variable sets of covariates to be recorded by different patients, adjust for missing data, allow high order interactions of covariates, and the adaptive allocation of each patient to treatment arms using the posterior predictive probability of which arm is best for each patient. The new method is demonstrated via extensive simulation studies.

Group Sequential Design Comparing Multiple Treatments to a Common Control
Jon Steingrimsson, Joshua Betz and Michael Rosenblum
Johns Hopkins University
jsteing@jhu.edu

We discuss a group sequential design comparing two treatments to a common control in two different sub-populations. The proposed
design leverages the correlation between the test statistics arising from that both treatments are compared to a common control. Furthermore, the design improves efficiency through step-down procedures and relies only on minimally sufficient statistics. Implementation of the design requires specifying several design parameters. We use optimization techniques to select the design parameters that minimize expected sample size subject to type one error and power constraints. The performance of the design is compared to more traditional designs using a recent clinical trial on cardiac resynchronization therapy.

Two Stage Drop the Loser Design in Clinical Trials with Response Adaptive Randomization
• Hongjian Zhu1, Jin Piao1, Jack Lee2 and Feifang Hu3
1 University of Texas Health Science Center at Houston
2 University of Texas MD Anderson Cancer Center
3 George Washington University
hongjian.zhu@uth.tmc.edu
Seamless phase II/III clinical trials have attracted increasing attention recently. They mainly use Bayesian response adaptive randomization (RAR) designs. There has been little research into seamless clinical trials using frequentist RAR designs because of the difficulty in performing valid statistical inference following this procedure and controlling the type I error rate. Well-designed frequentist RAR can target theoretically optimal allocation proportions, and they have explicit asymptotic results. We propose a family of RAR designs for seamless phase II/III clinical trials, study its asymptotic properties, and investigate how to control the type I error rate. The numerical studies demonstrate that the type I error rate can be well controlled with the proposed method and certain ethical and efficient objectives can be achieved using the proposed design.

Session 108: Topics on Applied Statistics (I)
Fitting Bayesian Spatial Survival Models Using R
• Haiming Zhou1 and Timothy Hanson2
1 Northern Illinois University
2 University of South Carolina
zhouh@niu.edu
A comprehensive, unified approach to modeling arbitrarily censored spatial survival data is presented for the three most commonly-used semiparametric models: proportional hazards, proportional odds, and accelerated failure time. Unlike many other approaches, all manner of censored survival times are simultaneously accommodated including uncensored, interval censored, current-status, left and right censored, and mixtures of these. Left-truncated data are also accommodated leading to models for time-dependent covariates. Both georeferenced (location observed exactly) and areally observed (location known up to a geographic unit such as a county) spatial locations are handled. Model fit is assessed with conditional Cox-Snell residual plots, and model choice is carried out via LPML and DIC. Baseline survival is modeled with a novel transformed Cox-Snell residual plots, and model choice is carried out via LPML spatial locations are handled. Model fit is assessed with conditional

A Robust Two-stage Design Identifying the Optimal Biological Dose for Phase I/II Clinical Trials
• Yong Zang1 and Jack Lee2
1 Indiana University
2 MD Anderson Cancer Center
zangy@iu.edu
We propose a robust two-stage design to identify the optimal biological dose for phase I/II clinical trials evaluating both toxicity and efficacy outcomes. In the first stage of dose finding, we use the Bayesian model averaging continual reassessment method to monitor the toxicity outcomes and adopt an isotonic regression method based on the efficacy outcomes to guide dose escalation. When the first stage ends, we use the Dirichlet-multinomial distribution to jointly model the toxicity and efficacy outcomes and pick the candidate doses based on a three-dimensional volume ratio. The selected candidate doses are then seamlessly advanced to the second stage for dose validation. Both toxicity and efficacy outcomes are continuously monitored so that any overly toxic and/or less efficacious dose can be dropped from the study as the trial continues. When the phase I/II trial ends, we select the optimal biological dose as the dose obtaining the minimal value of the volume ratio within the candidate set. An advantage of the proposed design is that it does not impose a monotonically increasing assumption on the shape of the dose-efficacy curve. Simulation results show that the proposed design has desirable operating characteristics across different shapes of the underlying true dose-toxicity and dose-efficacy curves.

A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies
• Jingjing Yang1, Lars Fritsche2, Xiang Zhou1 and Goncalo Abecasis1
1 University of Michigan
2 Norwegian University of Science and Technology
yijingj@gmail.com
Although genome-wide association studies (GWASs) have identified many complex loci, most of which reside in noncoding regions and have unknown biological functions. Integrative analysis that incorporates known functional information into GWAS can help elucidate the underlying biological mechanisms and prioritize causal-variants. We develop a flexible Bayesian variable selection model with efficient computational techniques for such integrative analysis. Different from previous approaches, our method models the effect-size distribution and probability of causality for variants with different annotations and jointly models genome-wide variants to account for linkage disequilibrium (LD), thus prioritizing associations based on the quantification of the annotations and allowing for multiple causal-variants per locus. Our method dramatically improves both computational speed and posterior sampling convergence by taking advantage of the block-wise LD structures in human genomes. In simulations, our method accurately quantifies the functional enrichment and performs more powerful for identifying true causal-variants than alternative methods, where the power gain is especially apparent when multiple causal-variants in LD reside in the same locus. We applied our method to an in-depth GWAS of age-related macular degeneration with 33,976 individuals and 9,857,286 variants. We find the strongest enrichment for causality among non-synonymous variants (54x more likely to be causal, 1.4x larger effect-sizes) and variants in active promoter (7.8x more likely, 1.4x larger effect-sizes), as well as identify 5 extra potential loci in addition to the 32 known AMD risk loci. In conclusion, our method is shown to efficiently integrate functional information in GWASs, helping identify causal-variants and underlying biology.

A Simple Method for Bayesian Variable Selection Based on Parameter Estimates, with Application to Me
• Grace Yoon, Wenxin Jiang and Lei Liu
Northwestern University
graceyoon@u.northwestern.edu
Title: A simple method for Bayesian variable selection based on parameter estimates, with application to medical cost data
Several unique statistical issues associated with health care cost, such as high proportion of zeros, heteroscedasticity and severe skewness, make statistical analysis challenging. Transformation of response variable has been alleviating those difficulties but has the main drawback that it is complicated to get the mean response in the original scale. To avoid these issues, we propose spike-or-slab priors for Bayesian variable selection based on frequentist parameter estimates. By ranking Z-statistics, the scope of model searching can be reduced to achieve computing efficiency. We apply this method to Medical Expenditure Panel Survey (MEPS) dataset.

Statistical Considerations in Using Meta-analysis for Regulatory Decision Making for Medical Devices
Qin Li
FDA Center for Devices and Radiological Health
qin.li@fda.hhs.gov
A well-designed meta-analysis may create an opportunity to combine and integrate comparable studies and trials, and the results from such analysis may be used to support assessment of clinical validity in regulating the medical devices when the studies are well conducted. However, there are also many challenges. For example, Meta-analysis may be subject to biases such as publication bias, selection bias, etc. Heterogeneity in the study population, study design and study conduct, etc. can create difficulties in making statistical inference and result interpretation. The quality assessment of selected publications in meta-analysis such as blinding, missing data, etc. is crucial in the evaluation. This talk will discuss these thoughts in using meta-analysis for regulatory decision making for medical devices.

Parallelized Adaptive Rejection Metropolis Sampling on the Utility of Graphics Cards
Yuefeng Wu
University of Missouri St. Louis
w96123051@gmail.com
Adaptive rejection Metropolis sampling (ARMS) is one of the popular tools for solving many modern-day statistical, optimization and computational problems. As it depends on the Markovian property in its Metropolis steps, a major limitation is the inherently sequential nature. The author proposes a nature generalization of the ARMS algorithm that allows for parallelization. The parallelized ARMS significantly decreases the variance of any estimator derived from the simulated samples, at a zero computing cost, since the improvements are based on a fixed number of target density evaluations that are calculated parallel. The proposed algorithm still holds the Markovian convergence properties as its single chain version due to the Rao-Blackwell principles it based on. Usually, parallel computing relies on clusters which are much more expensive and somehow harder to use than desktop computers. The author implements the parallelized algorithm on a lower end general purpose graphic process unity (GP GPU) on a cost of a fraction of a normal desktop itself.

Clinical Response Prediction with Logistical Model Adjusted by Tree Regression
Dion Chen
Janssen R&D, LLC
dchen97@its.jnj.com
In developing personalized medicine, it is important to separates patients into different groups with medical decisions to the individual patient based on their predicted responses or risk of disease. Biomarkers play an important role in developing personalized medicine. It is very interesting topic to predict the probability of patient clinical response after treatment period, with baseline biomarker. This paper introduces a new method of using a tree regression to improve logistic model for clinical response prediction based on baseline biomarker levels. The numerical results show that the linear logistic model can be significantly improved by a tree regression on the residuals. A simulation study indicates that the improvements could be significant with a moderate sample size. Although the classification problem of binary responses is discussed in this research, the idea is easy to extend to the classification of multinomial responses.

Keywords: personalized medicine; biomarker; tree regression; logistic model; classification

Characterizing Spatial Dependence on Stream Networks — Bayesian Hierarchical Model Approximation
Yingying Liu and Kate Cowles
The University of Iowa

As an extension of pairwise meta-analysis of two treatments, network meta-analysis has recently attracted many researchers in evidence-based medicine because it simultaneously synthesizes both direct and indirect evidence from multiple treatments and thus facilitates better decision making. The Lu/Ades Bayesian hierarchical model is a popular method to implement network meta-analysis, and it is generally considered more powerful than conventional pairwise meta-analysis, leading to more accurate effect estimates with narrower confidence intervals. However, the improvement of effect estimates produced by Lu/Ades network meta-analysis has never been studied theoretically. In this talk, we show that such improvement depends highly on evidence cycles in the treatment network. Specifically, Lu/Ades network meta-analysis produces posterior distributions identical to separate pairwise meta-analyses for all treatment comparisons when a treatment network does not contain cycles. Even in a general network with cycles, treatment comparisons that are not contained in any cycles do not benefit from Lu/Ades network meta-analysis. Simulations and a case study are used to illustrate the equivalence of Lu/Ades network meta-analysis and pairwise meta-analysis in certain networks.
Hock Peng Chan

**Multi-Sequence Segmentation Using Score Tests**

We propose local segmentation of multiple sequences sharing a National University of Singapore hypothesis. Moreover, we also address the problem of detecting ance and fourth order joint cumulants are constant under the null to previous work, our approach does not require that the mean, vari- correlation structures under less restrictive assumptions. In contrast the correlation. This paper develops change point analysis for the formed, it is not necessarily clear that the mean, marginal variance under the null hypothesis of no change-point”, which is crucial for understanding the influence of covariates on the response variable, especially in the presence of complex environmental factors. The methods developed in this paper can be applied to a wide range of environmental studies, including climate change, air pollution, and biodiversity conservation. Key Words: stream network, spatial dependence, Bayesian Hierarchical Model, Conditional Autoregressive Model

**Graphical Horseshoe Model for Inverse Covariance Estimation**

#### Purdue University

Yunfan Li, Anindya Bhadra and Bruce Craig

We develop a new estimator of the inverse covariance matrix for multivariate normal distributed data using the horseshoe prior. The proposed graphical horseshoe estimator resolves some of the common problems with other popular estimators, such as graphical lasso and graphical SCAD. The most prominent benefit is that the graphical horseshoe provides estimates with small information divergence from the true sampling distribution in sparse high dimensional problems, when alternative methods fail. We also provide a full Gibbs sampler for our estimator. Comparisons with alternative approaches demonstrate that the graphical horseshoe offers an attractive alternative to existing techniques, in terms of both statistical and computational performances.

**Session 110: Advances in change-point analysis**

**Change Point Analysis of Correlation in Non-stationary Time Series**

Holger Dette$^1$, Weichi Wu$^2$ and Zhou Zhou$^3$

$^1$Ruhr-University Bochum  
$^2$University College London  
$^3$University of Toronto

Zhou@utstat.toronto.edu

A restrictive assumption in change point analysis is “stationarity under the null hypothesis of no change-point”, which is crucial for asymptotic theory but not very realistic from a practical point of view. For example, if change point analysis for correlations is per- formed, it is not necessarily clear that the mean, marginal variance or higher order moments are constant, even if there is no change in the correlation. This paper develops change point analysis for the correlation structures under less restrictive assumptions. In contrast to previous work, our approach does not require that the mean, variance and fourth order joint cumulants are constant under the null hypothesis. Moreover, we also address the problem of detecting relevant change points.

**Multi-Sequence Segmentation Using Score Tests**

Hock Peng Chan

National University of Singapore

We propose local segmentation of multiple sequences sharing a common time- or location-index, building upon the single sequence local segmentation methods of Niu and Zhang (2012) and Fang, Li and Siegmund (2016). We show that multi-sequence local segmentation estimates change-points consistently, and that on both simulated and real datasets, signals are well-detected. We show that on a recent allele-specific copy number study involving multiple cancer patients, the simultaneous segmentations of the DNA sequences of multiple patients provide information beyond that obtained by segmentation of the sequences one at a time.

**Computationally Efficient Methods for Multivariate Change-point Analysis with Industrial Application**

Idris Eckley

Lancaster University

i.eckley@lancaster.ac.uk

In recent years there has been a surge of interest in the development of new methods for tackling change point problems. Much of this work has focused on the analysis of univariate time series, with contributions in this area having an impact in various disciplines, including climate science and genomics. However with the increased use of low-cost data collection methods, there is a growing demand to be able to efficiently and accurately identify changes in structure within highly multivariate time series. In moving to this more com- plex setting, the problem of detecting change points becomes more subtle and computationally intricate. This talk will outline some recent developments in this area that have been inspired by collabora- tions with industrial partners.

**DNA Copy Number Profiling Using Single-cell Sequencing**

Xuefeng Wang

H. Lee Moffitt Cancer Center & Research Institute

xuefeng.wang@moffitt.org

Currently, there is a lack of method and software for detecting copy number variations and constructing copy number profile for the whole genome from single-cell DNA sequencing data, which are often of low coverage and high technical noises. Here we intro- duce a new toolkit, SCNV, which features an efficient bin-free segmentation approach and provides the highest resolution possible for breakpoint detection and the subsequent copy number calling. SCNV can auto-tune parameters based on a set of normal cells from the same batch to adjust for the technical noise level of the data, facilitating its application to data gathered from different platforms and different studies.

**Session 111: The challenges of non-constant hazard ratio: delayed treatment effect, treatment dilution and treatment crossover**

**Statistical Issues with Clinical Trials where Proportional Haz- ards Assumption is Violated**

Xiaofei Wang and Stephen George

Duke University School of Medicine

xiaofei.wang@duke.edu

In designing a clinical trial for comparing two or more treatments with respect to time to event endpoint, a proportional hazards as- sumption is commonly made. However, in cancer immunotherapy trials, the immunotherapy drug often takes longer time to take effect than other therapies and causes the time-varying treatment ef- fect. Also, when overall survival is of primary interest, patients may pass through various disease states prior to death and because of this may receive treatments other than originally assigned. In these situations, the proportional hazards assumption will not hold. In the first part of the talk, we will emphasize that a simple but intuitive multi-state model allowing for progression, death before
Regression-Based Imputation Method

Comparison of the RMST with the Hazard Ratio in Superiority Trials with a Time-to-Event Endpoint

Bo Huang
Pfizer Inc.
bo.huang@pfizer.com

In a randomized clinical trial with a time-to-event endpoint, one primary objective is to quantify or measure the relative difference between the survival curves of the randomized arms, which is routinely characterized by a constant hazard ratio (HR) from the Cox proportional hazards model, under the assumption that the ratio of the two hazard functions is constant over time. With the emergence of novel therapies such as targeted therapies and immunotherapies exhibiting distinct mechanisms of action compared to traditional treatments, departure from the proportional hazard assumption in clinical trials with a time-to-event endpoint is increasingly common. In these situations, the HR may not be a valid statistical measurement of treatment effect and the corresponding log-rank test may no longer be the most powerful statistical test. We give an overview of alternative endpoints and analysis methods. In particular, we conduct a simulation study to evaluate the performance and operating characteristics of the Restricted Mean Survival Time (RMST)-based inference and against the hazard ratio based inference, under various scenarios and design parameter setups.

Adjusting for Subsequent Therapy in Oncology Trials: A Regression-Based Imputation Method

Tommy Fu1, Alan Wu2 and Xiaolong Lao2
1Celgene
tommyfu@gmail.com
2Celgene

In oncology trials, patients often receive subsequent therapies post study assigned treatments which confound the treatment effect on a long-term endpoint. A regression-based imputation procedure is proposed to mediate the confounding effect and to allow inferences on the “true” treatment effect. The characteristics of the procedure is evaluated by simulations. The proposed method has been applied to a randomized, phase 3 oncology trial to evaluate novel therapy over standard of care in elderly AML patients which ultimately led to regulatory approval.

Design and monitoring of survival trials with delayed treatment effect and treatment crossover

Xiaodong Lao1, Xuezhou Mao1, Xun Chen1, Junshan Qiu2, Steven Bai2 and Hu Quan1
1Sanofi
2FDA
xiaodong.luo@sanofi.com

Survival trials are often facing a lot of challenges in study design and monitoring when the proportional hazards assumption cannot be reasonably assumed. The common reasons for the non-proportional HR are delayed treatment effect and treatment crossover. A delayed treatment effect means the treatment will not have effect until enough exposure has been achieved therefore the HR is 1 in the beginning and becomes smaller at a later time. Treatment crossover means the study participants may cross from the assigned treatment to the other treatment arm therefore reducing the treatment effect. In this talk, we are presenting a design and monitoring tool to account for these complex situations.

Session 112: Adaptive Design

Design and Monitoring of Survival Trials with Delayed Treatment Effect and Treatment Crossover

Meiyu Shen and Lixin (Leo) Xu
CDER, FDA
meiyu.shen@fda.hhs.gov

During the biotechnology product development, the analytical instrumentation and methodology are often carefully selected based on the intended purpose and the scope of the analytical method. After an analytical method is successfully validated and implemented, the updates of the method with the associated standard operating procedure (SOP) will be conducted throughout the lifecycle of the product. Frequently, a transfer of a validated analytical method from an original laboratory (sending laboratory) to a new testing laboratory (receiving laboratory) is a common practice during the life-cycle management of analytical methods. Since the analytical method to be transferred has been already thoroughly evaluated and fully validated for its intended purpose at the sending laboratory, the main purpose of method transfer studies is usually the evaluation to determine if the two laboratories generate comparable results across the parameter ranges of interest, and to assure that the method after the transfer is still suitable for its intended use. Inconsistent advice is often seen regarding testing materials, statistical methods, and acceptance criteria in literature for the method transfer. Furthermore, there is no detailed recommendation for the design and analyses for method transfer studies in the regulatory guidance from US Food and Drug Administration and other regulatory authorities. This presents a significant challenge in the area of analytical method transfer for biotechnology products. We are proposing a new statistical methodology to tackle this challenge, i.e., the development of a design and a statistical equivalence analysis in which the determination of the margin for the mean equivalence between the data obtained in sending and receiving laboratories is a function of the sample size.

Dilemma on Conditional Power-Based Adaptive Sample Size Re-estimation

Xiaoyu Cai, Yi Trong, Meiyu Shen and Yu-Ting Weng
nkcaixiaoyu@gmail.com

Conditional power method was widely discussed in the literature on adaptive un-blinded sample size re-estimation (SSR) designs. The common application of conditional power method can be divided into three categories: (a) Futility assessment (b) Decision on sample size adjustment and (c) Decision on sample size increment. The conditional power based methods have advantages of controlling type I error rate and maintaining conditional power of the final test at a desired level. However, like most of other designs, it is not a universal optimal design, even among different approaches of adaptive designs. Moreover, the performance of conditional power method highly relies on some operational factors and sample distribution parameters such as the information fraction, the true treatment and so on. The problem is, except the basic statistical operating characteristics (type I error rate, power and so on), we have plenty of
different criteria to compare different designs but hardly identify a most important criteria. For example, we may get better estimation of conditional power but lost efficiency under the same information fraction. Several problems may make it a dilemma whether or not it is worthwhile to use conditional power based adaptive design in the clinical trials. In this study, we will point out some of these potential problems.

Utilizing Seamless Adaptive Designs for NASH Clinical Trials
‡Yeh-Fong Chen, Feiran Jiao and George Kordzakhia
US Food and Drug Administration
yehfong.chen@fda.hhs.gov

The prevalence of non-alcoholic fatty liver disease including non-alcoholic steatohepatitis (NASH) is increasing worldwide. NASH is the second most common indication for liver transplantation and is expected to be the leading indication by 2020. In light of increasing prevalence and burden of disease, it is imperative to develop therapeutic strategies for patients with NASH. Although it is feasible to conduct clinical trials with liver transplantation or death as the clinical outcome endpoint, it may take 10 to 20 years for NASH patients to develop cirrhosis or other liver-related morbidity and mortality. In addition, even with accelerated regulatory approval based on surrogate endpoints, studies need to be at least one year long to demonstrate clinically meaningful effectiveness in study drugs. Seamless adaptive clinical designs that “roll over” patients from earlier phases can be adopted to shorten the duration of new drug development programs for NASH. To address the complex multiplicity problems resulting from multiple doses and multiple looks, these types of designs require flexible and powerful (or efficient) multiple testing procedures. Once the drug is partially approved (with accelerated approval), additional challenges include how to ensure later studies, such as phase 4 studies, can be properly conducted and which controls groups are feasible. In this presentation, we will focus on the statistical considerations and the application of the seamless design in NASH clinical trials. Thorough exploration, including different methods for efficiently controlling study-wise type I error rate as well as the potential use of historical controls, will be illustrated and discussed.

Session 113: Recent Development and Applications of Statistical Methodology in Rare Diseases

Randomized phase II trials vs. single arm phase II trials with a registry control in rare diseases
‡Brent Logan, Ruta Brazdaitis and Raphael Fraser
Medical College of Wisconsin
clogan@mcw.edu

Randomization in phase II trial designs has been recognized as a means to reduce bias associated with single arm phase II trials being compared with historical data. However, in rare disease settings, randomized phase II trials are more difficult due to the rarity of patients available for study. Furthermore, there may be rare disease outcome registries available which can provide a control group for comparison which is large enough to allow for matching or covariate adjustment. While using such a registry control group does not guarantee an unbiased comparison, it may be more efficient (lower variability) and cost effective than randomized phase II trials in terms of requiring fewer patients on trial to detect the same treatment effect. It is important to study this strategy and understand the tradeoffs in bias/efficiency for realistic rare disease trials and endpoints. We demonstrate an investigation of these tradeoffs for potential trials examining several common outcomes after a hematopoietic cell transplant for hematologic malignancies. We simulate trials of each type (randomized phase II trials vs. single arm trials with registry control group) from real populations defined by existing completed randomized phase III clinical trials conducted by the Blood and Marrow Transplant Clinical Trials Network (BMTCTN), as well as available registry data from the Center for International Blood and Marrow Transplant Research (CIBMTR).

What Constitute Scientific Evidence? The Similarity Principle
Mark Chang
Veristat
mark.chang@veristat.com

Rear disease trial, personalized medicine, and Simpson Paradox Challenges in clinical trial for rear disease are discussed in terms of single-arm and two-arm designs. The issue trigger the fundamental question: what should constitute scientific evidence.

Adaptive Multi-Stage Clinical Trial Design for Binary Endpoint in Rare Disease
‡Zhaowei Hua
1 Takeda Pharmaceutical International Co.
2 University of Illinois at Urbana and Champaign
zhaowei.hua@takeda.com

Development of therapeutic agents in the field of rare disorders is challenged by the limited number of patients eligible to be studied. A small sample size could lead to invalidity of asymptotics of commonly used testing statistics such as chi-squared statistic, and leave power of conventional randomized clinical trial not adequately enough. The article proposes an adaptive multi-stage clinical trial design for binary endpoint in rare diseases. It presents an exact unconditional test statistic to generally control type I error when sample size is really small while not sacrificing power. Adaptive randomization is used to increase the chance for patients to be allocated to more effective treatment as well as increase power. The methods are illustrated using simulations with clinical trials of sample size as small as 20.

A Bayesian prediction model between a biomarker and the clinical endpoint for dichotomous variables
‡Yang Song1, Zhiwei Jiang2, Qiong Shou2, Jielai Xia2 and William Wang2
1 Vertex Pharmaceuticals Inc., 2 Merck Research Labs, 3 Fourth Military Medical University of China
yang_song@vrtx.com

Motivated by a regional vaccine study where the clinical endpoint only provides a very small number of events to allow confirmatory statistical inference, we propose a Bayesian meta-analytic approach to build and evaluate an association model between an early biomarker endpoint and the late clinical endpoint from historical clinical trials. Extensive simulation studies have been conducted to assess robustness and applicability of the model. When this model is applicable, statistical inference could be conducted in the early biomarker endpoint with more information. Such an approach could be considered in rare disease drug development where small sample size due to low disease prevalence often poses challenges to adequately power clinical trials.
Session 114: Statistical Genetics & Genomics (or Computational Biology)

**GEE-based SNP set association testing for longitudinally-measured continuous and discrete traits**
Zuoheng Wang
Yale University
zuoheng.wang@yale.edu

Recent developments in high-throughput sequencing technologies have made it possible to search for both rare and common genetic variants associated with complex diseases. Many phenotypes in health studies are measured at multiple time points. The rich information on repeated measurements on each subject not only provides a more accurate assessment of disease condition, but also allows us to explore the genetic influence on disease onset and progression. However, most association tests mainly focus on a single point. To address this limitation, we propose a generalized estimating equations (GEE)-based kernel association test to assess association between a SNP set and longitudinal data, which extends the GEE method for a single SNP to a set of SNPs. The proposed method allows for both longitudinally-measured continuous and discrete traits, where the within-subject correlation is taken into account by using the empirical covariance estimator. In simulation studies, we evaluate the performance of the new method, and demonstrate that it has improved power, by making full use of multiple measurements, as comparing to previously proposed tests on a single measurement or average measurements for each subject. We illustrate the new method in two real data sets.

**Weighted score method for twin imaging study detects genetic and environmental effects in Human conn**
Yimei Li1, Shenghua Mao1 and Hongtu Zhu2
1St. Jude Children’s Research Hospital
2MD anderson cancer center
yimei.li@stjude.org

Twin imaging studies have been valuable for understanding the relative contribution of the environment and genes on brain structures and their functions. Conventional analyses of twin imaging data include three sequential steps: spatially smoothing imaging data, independently fitting a structural equation model at each voxel, and finally correcting for multiple comparisons. However, conventional analyses are limited due to the same amount of smoothing throughout the whole image, the arbitrary choice of smoothing extent, and the decreased power in detecting environmental and genetic effects introduced by smoothing raw images. The goal of this paper is to develop a weighted score test method for spatial analysis of twin neuroimaging and behavioral data. This method can be used to establish the relationship between twin imaging data and a set of covariates of interest, such as age and gender as well as disentangle the environmental and genetic influences on brain structures and their functions. Simulation studies show that our weighted score method significantly outperforms conventional analyses of twin imaging data. Finally, we use our method to detect statistically significant effects of genetic and environmental variations on white matter structures for twin participants in Human connectome project.

How to define negative data in tumor heterogeneity study?
Sujana Ganapathi1, Mitsuko Murakami1, Li Yan1, Jianmin Wang1, Changxing Ma2, Song Liu3 and Lei Wei1
1Roswell Park Cancer Institute
2University at Buffalo
lei.wei1@roswellpark.org

Tumor heterogeneity plays an important role in molecular diagnosis, drug resistance and cancer recurrence. Characterizing tumor heterogeneity on the molecular level requires comparison of mutational profiles between different tumor samples. One challenge lies in defining negative status - when a mutation is not detected in a sample, it can be either caused by absence of the mutation or insufficient coverage. However, there is no definitive way to determine a minimum coverage needed because it depends on the mutation’s relative abundance. Currently most tumor heterogeneity studies use an overall coverage cutoff for all mutations. This can lead to misclassification between negative and unknown (low coverage) statuses. We hypothesized that negative statuses can be more accurately defined on individual mutation’s basis. Using a unique test data set of single-cell dual-platform sequencing, our proposed new method achieved better cross-platform concordance than any traditional single coverage cutoff method. Our results provide a more precise approach for distinguishing negative from unknown statuses for any tumor heterogeneity study.

Integrated Statistical Inference on Genomic Handles of Traits (InSIGHT)
Cheng Cheng and Lei Shi
St. Jude Children’s Research Hospital
cheng.cheng@stjude.org

The genomic biochemical process underlying a complex trait often involves a large number and several types of molecules derived from the genome. The advancement of biotechnology has enabled researchers to measure genomic and epigenomic elements (DNA sequence variations, methylation levels, microRNA expressions), as well as gene (and protein) expressions, on genome-wide scale and often in relatively large cohorts of subjects. The technological advancement provides opportunities to gain more comprehensive understanding of the genomic biological process underlying complex traits, yet raises a challenge for statistical inference in terms of how information from different types of molecules can be “fused” to enhance the power of discovering genomic regions (especially genes) responsible for the trait variations. In this talk we present one component of InSIGHT – the integrated inference of all measured cis elements around genes for their associations with a quantitative trait. We will discuss a novel procedure called truncated aggregation of P values (TAP) test, with a description of the methodology, a real data example, and if time permits, some simulation results.

Session 115: Recent Advances in Dose Finding studies

**Quantitative consideration in dose-finding study designs**
Huilin Hu and Dong Xi
Novartis Pharmaceutical Corp
huiulin.hu@novartis.com

Dose-finding studies not adequately designed to support the characterization of dose response and proper selection of dose(s) have been considered as one of the root causes for high failure rate in the Phase III program. Although literatures are abundant to discuss the issues and possible solutions to improve the design, designing a dose finding study is still considered one of the most challenging steps in the drug development. The level of complexity involved varies depending on the availability of knowledge (such as dose-range, study population, endpoint, etc) at the design stage. There are many essential components to consider when designing a dose-finding study, but in this talk we focus on the aspect of sample size determination. We consider it important to size a study according to what it meant to support in the development decision-making beyond the detection...
of dose-response signal and precision of dose-response estimates. The study objectives in a dose finding study are often formulated in somewhat vague terms, such as “to characterize the dose-response relationship”. To better evaluate and optimize the trial design, it is important to have the objectives translated into clear quantitative and actionable statements and especially to ensure a close connection to the development decision-making, e.g. whether or not the development should be terminated early, whether to GO/NO-GO to Phase III, and which dose(s) (and/or regimen) to bring forward to confirmatory studies, etc. The idea will be illustrated by case studies.

Dose Response Models for Longitudinal Data

CHYIHUNG HSU
Janssen Research & Development, Raritan, NJ, USA

Janssen Research & Development, Raritan, NJ, USA

Dose-regimen selection for confirmatory trials and characterization of dose-response relationship are unarguably among the most important and difficult tasks in clinical drug development. Inadequate understanding of the dose-response relationship is believed to be one of the key drivers of the high attrition rate in Phase III. In this presentation, longitudinal data models will be leveraged to enhance the efficiency of estimating dose-response relationship. Simulations will be performed to evaluate and quantify additional efficacy gain from longitudinal modeling. Limitations of this approach will also be discussed.

Assessing the similarity of dose response and target doses in two non-overlapping subgroups

Frank Bretz
Novartis
frank.bretz@novartis.com

We consider two problems of increasing importance in clinical dose finding studies. First, we assess the similarity of two non-linear regression models for two non-overlapping subgroups of patients over a restricted covariate space. To this end, we derive a confidence interval for the maximum difference between the two given models. If this confidence interval excludes the equivalence margins, similarity of dose response can be claimed. Second, we address the problem of demonstrating the similarity of two target doses for two non-overlapping subgroups, using again a confidence interval based approach. We illustrate the proposed methods with a real case study and investigate their operating characteristics via simulation.

Session 116: Partial Identification: Seeking reasoned evidence from difficult data

Bayesian inference for partially identified models: Exploring the limits of limited data

Paul Gustafson
University of British Columbia
pgustaf@stat.ubc.ca

Intuitively, partial identification arises when the true value of the target parameter cannot be recovered in the large-sample limit, but some values of the target can be ruled out. Models for coarsened data, for instance, can give rise to only partial identification. There is no conceptual problem in turning the Bayesian crank in a partially identified setting. There is, however, a practical need to understand the utility of such inference. This talk explores this issue. One emphasis is on untangling when a parameter governing the data coarsening mechanism is purely a sensitivity parameter, as opposed to being a parameter that can be informed by data.

Estimating the prevalence of accounting misconduct: A semiparametric Bayesian approach

Richard Hahn1, Jared Murray2 and Ioanna Manolopoulou3
1University of Chicago
2Carnegie Mellon University
3University College of London
jsmurray@stat.cmu.edu

The United States Securities and Exchange Commission (SEC) releases public notices called Accounting and Auditing Enforcement Releases (AAERs). Informally, AAERs comprise a list of publicly traded firms that the SEC has cited for misconduct in one form or another. From the AAER data it is straightforward to estimate the probability that a firm will get caught engaging in misconduct, but the more natural estimand is the probability that a firm engages in misconduct at all. The observable data are only partially informative about this quantity (providing a lower bound).

To make inference in this setting we propose a general method for using flexible Bayesian regression models to estimate a partially identified probability function. Our approach allows us to be selectively informative, using nonparametric prior distributions on quantities identifiable from the observed data and informative priors on quantities that are only partially identified. It also admits computationally efficient sensitivity analysis concerning the posterior impact of priors on the partially identified component of the regression model.

Partially Identified Treatment Effects for Generalizability

Wendy Chan
University of Pennsylvania
wechan@upenn.edu

Recent methods to improve generalizations from nonrandom samples typically involve assumptions such as the strong ignorability of sample selection, which is challenging to meet in practice. Although researchers acknowledge the difficulty in meeting this assumption, point estimates are still provided and used without considering alternative assumptions. We compare the point identifying assumption of strong ignorability of sample selection with two alternative assumptions bounded sample variation and monotone treatment response that partially identify the parameter of interest, yielding interval estimates. Additionally, we explore the role that population data frames play in contributing identifying power for the interval estimates. We situate the comparison around causal generalization with nonrandom samples by applying the assumptions to a cluster randomized trial in education. Bounds on the population average treatment effect are derived under the alternative assumptions and the case when no assumptions are made on the data. While comparing the bounds, we discuss the plausibility of each alternative assumption and the practical tradeoffs. We highlight the importance of thoughtfully considering the role that assumptions play in causal generalization by illustrating the differences in inferences from different assumptions.

Session 117: High dimensional multivariate analysis and its applications

Two-Sample Tests for Sparse High Dimensional Multinomial Distributions

Amanda Plunkett1 and Jiunyong Park2
1Department of Defense, U.S.A.
2University of Maryland Baltimore County
junpark@umbc.edu

Abstracts
We consider testing the equality of probability vectors of two independent multinomial distributions. The classical chi-square test may have some drawbacks in this case since some of the cell counts may not be large enough. We propose new type of tests and show their asymptotic normality and the asymptotic power functions. To compare with existing tests, we provide numerical studies including simulations and some applications.

Change-point detection for locally dependent data
Hao Chen
University of California, Davis
hchen@ucdavis.edu

Local dependence is common in multivariate and object data sequences. We consider the testing and estimation of change-points in such sequences. A new way of permutation, circular block permutation with a randomized starting point, is proposed and studied for a scan statistic utilizing graphs representing the similarity between observations. The proposed permutation approach could correctly address for local dependence and make it possible the theoretical treatments for the non-parametric graph-based scan statistic for locally dependent data. We derive accurate analytic approximations to the significance of graph-based scan statistics under the circular block permutation framework, facilitating its application to locally dependent multivariate or object data sequences.

Test for the mean matrix in a Growth Curve model for high dimensions
Muni Srivastava
University of Toronto
srivasta@utstat.toronto.edu

We consider the problem of estimating and testing a general linear hypothesis in a general multivariate linear model, the so-called Growth Curve Model, when the p x N observation matrix is normally distributed with an unknown covariance matrix. The maximum likelihood estimator (MLE) for the mean is a weighted estimator with the inverse of the sample covariance matrix which is unstable for large p close to N and singular for p larger than N. We modify the MLE to an unweighted estimator and propose a new test which we compare with the previous likelihood ratio test (LRT) based on the weighted estimator, i.e., the MLE. We show that the new test based on the unweighted estimator performance better than the LRT. For the high-dimensional case, when p > N, we construct two new tests based on the trace of the variation matrices due to the hypothesis (between sum of squares) and the error (within sum of squares). To compare the performance of these four tests we compute the attained significance level (ASL) and the empirical power.

New Insights in High Dimensional Tests with Applications to Genetic and Genomic Studies
Lingzhou Xue1 and Danning Li2
1Penn State University
2Jilin University
lxue@psu.edu

In this talk, I will discuss some new insights in hypothesis tests for analysis of high-dimensional data, which are motivated by genetic and genomic studies. In the current literature, two sets of test statistics are commonly used for various high-dimensional tests: 1) using extreme-value form statistics to test against sparse alternatives, and 2) using quadratic form statistics to test against dense alternatives. However, quadratic form statistics suffer from low power against sparse alternatives, and extreme-value form statistics suffer from low power against dense alternatives with small disturbances and may have size distortions due to its slow convergence. For real-world applications, it is important to derive powerful testing procedures against more general alternatives. Based on their joint limiting laws, we introduce new testing procedures to boost the power against more general alternatives and retain the correct asymptotic size. Under the high-dimensional setting, we derive the closed-form limiting null distributions, and obtain their explicit rates of uniform convergence. We demonstrate the performance of our proposed test statistics in numerical studies.

Session 118: Recent Developments in Graphical Models and Network Analysis

Low-rank Tensor Recovery via Cubic-Sketching
Botao Hao1, Anru Zhang2 and Guang Cheng2
1Purdue University
2University of Wisconsin Madison
haobotao000@gmail.com

In this presentation, we propose a general framework for recovering sparse and low-rank tensors through rank-one cubic sketching. Two real-world applications are considered: one on high-dimensional interaction models; another on compressed image transmission. A block-wise thresholded gradient decent algorithm is proposed for stable recovery in both noiseless and noisy cases. Both upper bound and lower bound for the estimation accuracy are obtained over a large class of low-rank tensors, demonstrating the optimality of the proposed procedure.

Estimation of Gaussian Graphical Model from Data with Dependent Noise Structure
Yuying Xie1, Yufeng Liu2 and William Valdar2
1Michigan State University
2University of North Carolina at Chapel Hill
xyy@stt.msu.edu

Gaussian graphical models (GGMs) are commonly used to represent conditional dependence among random variables. A notable drawback of existing methods for estimating GGMs is that they ignore the existence of measurement error, which is both common and varied in biological data. It has been shown that RNA-seq could reliably quantify expression of only 30% of the genes with a relative error less than 20%, and many factors could introduce measurement error to microarray and RNA-Seq results including RNA degradation, PCR amplification efficiency, and RNA secondary structure. In this paper, we model the observed random vector as the sum of the true random vector plus random noise with dependent structure. In this setting, we show that the underlying GGM is not identifiable. To address this issue, we proposed a new experimental design using technical replicates and developed a new methodology to estimate the sparse GGM while taking account the underlying measurement errors. Additionally, the estimation consistency and selection sparsity of proposed method were established. Numerical results suggested that the proposed algorithm is superior to existing methods in both estimation and selection accuracy.

Nonparametric Seeded Network Matching
Yuan Zhang
The Ohio State University
zhang.7824@osu.edu

Seeded network matching is a semi-supervised learning problem that attracts raising interests in recent years. The task is to pair up nodes in two networks, provided just a few correctly matched
node pairs. The problem is typical in the “cheap data, expensive manpower” era, and the topic finds a wide range of applications in engineering, social sciences, biomedical and biological studies and so on. Existing methods usually require modeling conditions such as that the two networks are perfectly matchable and generated from certain types of block models. In this paper, we propose a method that can match networks under more general model assumptions. Our method can handle both non-matchable nodes and many-to-one or one-to-many maps. We show both theoretical guarantee and competitive numerical performances of our method.

Session 119: Novel statistical methods for genetic data analysis

Empirical Likelihood Ratio Tests for Coefficients in High Dimensional Heteroscedastic Linear Models

Hongfang Wang 1, Ping-Shou Zhong 2 and Yuehua Cai 2
1 Indiana University-Purdue University Indianapolis
2 Michigan State University

hilwang@iupui.edu

This paper considers hypothesis testing problems for a low-dimensional coefficient vector in a high-dimensional linear model with heteroscedastic variance. Heteroscedasticity is a commonly observed phenomenon in many applications including finance and genomic studies. Several statistical inference procedures have been proposed for low-dimensional coefficients in a high-dimensional linear model with homoscedastic variance. However, existing procedures designed for homoscedastic variance are not applicable for models with heteroscedastic variance and the heteroscedasticity issue has been rarely investigated and studied. We propose a simple inference procedure based on empirical likelihood to overcome the heteroscedasticity issue. The proposed method is able to make valid inference even when the conditional variance of random error is an unknown function of the high-dimensional predictor. We apply our inference procedure to three recently proposed estimating equations of disease heterogeneity. Advanced analytical methods are in great need to account for the genetic heterogeneity of complex human diseases. However, detecting these disease-susceptibility rare variants remains a great challenge because of the heterogeneous nature and low frequency of rare variants. Multiple rare variants within the same gene can independently influence the disease (i.e., allelic heterogeneity), and rare variants in different genes can also be involved in related pathways underlying diseases (i.e., locus heterogeneity). Advanced analytical methods are in great need to account for the genetic heterogeneity of complex human diseases. In this talk, we introduce a family-based genetic random filed (FGRF) method for association analyses of sequencing data in family-based association studies. By utilizing information from family members, the proposed method is robust to population stratification and gains improved performance in presence of genetic heterogeneity. The proposed method is compared to other existing methods through simulation studies and real data applications for investigating the

Secondary traits - rare variants association analyses in case-control sequencing studies

Guolian Kang
St. Jude Children’s Research Hospital
guolian.kang@stjude.org

In many case-control designs of genome-wide association (GWAS) or next generation sequencing (NGS) studies, extensive data on secondary traits that may correlate and share the common genetic variants with the primary disease status are available. Investigating these secondary traits can provide critical insights into the disease etiology or pathology and enhance the GWAS or NGS results. Methods based on logistic regression (LG) were developed for this purpose. However, for the identification of rare variants (RVs) underlying common diseases, certain inadequacies in the LG models and algorithmic instability can cause severely inflated type I error probability and significant loss of power when the two traits are correlated and the RV is associated with the disease status, especially at a stringent significance level. To address this issue, we propose a novel set-valued (SV) model that models a binary trait by dichotomization of an underlying continuous variable and incorporates this into the genetic association model as a critical component. Extensive simulations and an analysis of 7 secondary traits in a GWAS of benign ethnic neutropenia show that the SV method consistently controls type I error well at stringent significance levels, has larger power than the LG-based methods, and is robust in performance to effect pattern of the genetic variant (risk or protective), rare or common variants, rare or common diseases, and trait distributions. Because of the SV method’s striking and profound advantage, we strongly recommend the SV method be employed instead of the LG-based methods for secondary traits analyses in case-control sequencing studies.

Random Field Modelling of Genetic Association in the presence of disease heterogeneity

Ming Li 1, Zihuai He 2 and Qing Lu 3
1 Indiana University at Bloomington
2 University of Michigan
3 Michigan State University

li498@indiana.edu

Emerging studies using next-generation sequencing technology hold great promise for the identification and fine mapping of novel genetic variants, especially rare variants, contributing to complex human diseases. However, detecting these disease-susceptibility rare variants remains a great challenge because of the heterogeneous nature and low frequency of rare variants. Multiple rare variants within the same gene can independently influence the disease (i.e., allelic heterogeneity), and rare variants in different genes can also be involved in related pathways underlying diseases (i.e., locus heterogeneity). Advanced analytical methods are in great need to account for the genetic heterogeneity of complex human diseases. In this talk, we introduce a family-based genetic random filed (FGRF) method for association analyses of sequencing data in family-based association studies. By utilizing information from family members, the proposed method is robust to population stratification and gains improved performance in presence of genetic heterogeneity. The proposed method is compared to other existing methods through simulation studies and real data applications for investigating the
Session 120: Complex Data Analysis: Methodologies and Applications

Bayesian sensitivity analysis for unmeasured confounding in causal mediation analysis

Lawrence McCandless
Simon Fraser University
lmcandl@sfu.ca

Unmeasured confounding is a serious concern in observational studies using large administrative databases. The massive sample size collapse p-values and standard errors to zero that are calculated from standard analytic adjustment. While this may delight health researchers who discover that everything is significant, it obscures the role of bias, including unmeasured confounding. The Bayesian approach to statistics provides an appealing way forward because uncertainty about bias can be funnelled into the analysis using prior distributions. The posterior distribution for model parameters incorporates uncertainty from bias in addition to the usual random error. In this work I present a Bayesian approach to explore sensitivity to unmeasured confounding in causal mediation analysis with confounding in the mediator-outcome relationship. The method is illustrated in a data example from social epidemiology using large administrative databases of electronic health records. I demonstrate that great care is needed in choosing the prior distribution in a non-identified context because it can have surprising unexpected influence on the analysis results.

Functional Principal Component Analysis for Longitudinal and Survival Data

Liangliang Wang and Jianghu Dong
Simon Fraser University
lwa68@sfu.ca

Our work is motivated by longitudinal clinical data and survival responses of kidney transplant recipients. We use a nonparametric method to jointly model longitudinal and survival data using functional principal components. Kidney function process is recorded as estimated glomerular filtration rates (GFR) at multiple time points post transplantation. The longitudinal trajectories of GFR are assumed to have the measurement error. These trajectories are represented by flexible basis functions, such as B-splines, and the model dimension is reduced by functional principal component analysis. We assess the relationship between the longitudinal process and survival responses by exploring several survival models. We propose a novel EM algorithm to estimate the parameters in the model.

Can we use machine learning with large administrative datasets? Should we?

Robert Platt
McGill University
robert.platt@mcgill.ca

In analyses of drug safety studies using large administrative datasets, confounding by indication is an important source of bias. The propensity score is often used to address these biases. When large numbers of covariates are available for analysis, methods such as the high-dimensional propensity score (hdPS) are often used; this method has been shown to have reasonable properties under a range of circumstances but has not been systematically evaluated. Standard propensity score methods may also introduce bias when relations between covariates and treatment are non-linear and the model is mis-specified. We compare the standard propensity score and the hdPS to a variety of machine-learning prediction algorithms to estimate the propensity score for use in inverse weighted estimation of a marginal structural Cox model; machine learning tools demonstrate small bias and mean square error.

Analysis of Longitudinal Data with Missing Observations or Measurement Error

Grace Yi1, Xianming Tang2 and Runze Li3
1University of Waterloo
2UNC Chapel Hill
3Penn State
yyi@uwaterloo.ca

In contrast to extensive attention on model selection for univariate data, research on correlated data remains relatively limited. Furthermore, in the presence of missing data and/or measurement error, standard methods would typically break down. To address these issues, we propose marginal methods that simultaneously carry out model selection and estimation for longitudinal data analysis. Our methods have a number of appealing features: the applicability is broad because the methods are developed for a unified framework with marginal generalized linear models; model assumptions are minimal in that no full distribution is required for the response process and the distribution of the mismeasured covariates is left unspecified; and the implementation is straightforward. To justify the proposed methods, we provide both theoretical properties and numerical assessments.

Session 121: Multiplicity in Clinical Trials

Multiple endpoints in clinical trials: the regulatory, statistical, and common sense perspectives

Dror Rom and Jaclyn McTague
Prosoft Clinical
d.rom@prosoftclinical.com

Confirmatory clinical trials are often required to produce robust efficacy and safety information. These requirements are achieved through strong effects (efficacy), or lack thereof (safety) typically by large sample sizes, multiple assessments (endpoints) of safety and efficacy parameters, different sub-groups, time-points, etc. In some therapeutic areas, as either a requirement, or as a desired feature, studies may involve a primary and an important secondary endpoint, or co-primary endpoints. It has been argued (for example, Alosh and Huque, 2010), that two endpoints that can fully characterize a treatment effect on their own, should produce results that are consistent with each other, otherwise, a favorable interpretation of the efficacy of a new treatment will be difficult to make. A common requirement in the scenario above is that the testing procedure leads to the conclusion that at least one endpoint, and hopefully both, produce statistically significant favorable treatment effects. Clinical and statistical considerations must be given to the selection of the co-primary endpoints and the testing scheme. When both co-primary endpoints are expected to trend favorably, it is prudent to select a test procedure that is powerful for such a setting. Test procedures that are not specifically designed to address this, for example, the Bonferroni procedure and some of its modifications, may be sub-optimal.

To this end, we examine the Alosh and Huque (2010) procedure that incorporates a consistency parameter into the testing scheme, and formulate a non-parametric flexible class of procedures that have strong control of the Familywise Error rate. We examine the type-I error and power characteristics of these procedures in comparisons to other procedure often used in the above settings.
A Gatekeeping Test in a Group Sequential Design with Multiple Interim Looks
Ajit Tamhane1, *Jiangtao Gou2, Christopher Jennison3, Cyrus Mehta1 and Teresa Carito4
1Northwestern University
2Hunter College of CUNY
3University of Bath
4Cytel Inc.

Glimm et al. (2010) and Tamhane et al. (2010) studied the problem of testing a primary and a secondary endpoint, subject to a gatekeeping constraint, using a group sequential design (GSD) with K = 2 looks. We extend the previous results to multiple (K > 2) looks.

Statistical Considerations in Un-Blinded Sample Size Re-Estimation in a Phase 3 Trial
♦ Xin Wang, Sheng Zhong, Yijie Zhou and Lu Cui
Abbvie
xin.wang@abbvie.com

Abstract: Un-blinded sample size re-estimation is an adaptive design method that re-estimates (mostly increases) sample size when the interim observed effect is promising, but not as large as the planned effect and thus the final analysis may not be adequately powered. The conditional power at interim analysis is usually used to determine if the sample size should be increased. We consider this design approach in the context of a Phase 3 trial, where it is desirable to demonstrate non-inferiority of study drug as compared to an active control. The fixed sample size design for non-inferiority (NI) comparison is under the assumption that the test drug will have the same efficacy as the active control. Due to lack of data to inform on this assumption, we propose un-blinded sample size re-estimation to mitigate the risk of a small shift from the assumed active control effect. Several innovative steps are taken when implementing un-blinded sample size re-estimation in this context. Firstly, we adapt the methodology to NI comparison setting while it was originally developed for superiority testing. Secondly, the study has 2:1 randomization ratio for the study drug vs. active control, therefore instead of the standard approach to increase across all study arms, it is potentially more efficient to increase only the active control arm. We conduct simulations to compare our approach to the traditional approach with proportional sample size increase in terms of type I error control and power boost, as well as operational challenges. Weighted and un-weighted test statistics are examined in the simulations.

Session 122: high dimensional modeling in medicine

Regularized Estimation in Sparse High-dimensional Multivariate Regression in DNA Methylation Study
♦ Lei Liu1, Haixiang Zhang2, Lifang Hou2, Wei Zhang1, Andrea Baccarelli3, Yinan Zheng1 and Grace Yoon3
1Northwestern University
2Tianjin University
3Columbia University
lei.liu@northwestern.edu

In this article, we consider variable selection for correlated high dimensional DNA methylation markers as multivariate outcomes. A novel weighted square-root LASSO procedure is proposed to estimate the regression coefficient matrix. A key feature of this method is tuning-insensitivity, which greatly simplifies the computation by obviating the cross validation for penalty parameter selection. A precision matrix which can be obtained via the constrained minimization method (Cai et al., 2011) is used to account for the within-subject correlation among correlated outcomes. Oracle inequalities of the regularized estimators are derived. The performance of our proposed methodology is illustrated via extensive simulation studies. We apply our method to study the relation between smoking and high dimensional DNA methylation markers in the Normative Aging Study (NAS).

Personalized Medicine for Survival Outcomes
Haoda Fu
Eli Lilly and Company
fuhaoada@gmail.com

ITR is a method to recommend treatment based on individual patient characteristics to maximize clinical benefit. During the past a few years, we have developed and published methods on this topic with various applications including comprehensive search algorithms, tree methods, benefit risk algorithm, multiple treatment & multiple ordinal treatment algorithms. In this talk, we propose a new ITR method to handle survival outcomes for multiple treatments. This new model enjoy the following practical and theoretical features
(1) Instead of fitting the data, our method directly search the optimal treatment policy which improve the efficiency (2) To adjust censoring, we propose a doubly robust estimator. Our method only requires either censoring model or survival model is correct, but not both. When both are correct, our method enjoys better efficiency (3) Our method handles multiple treatments with intuitive geometry explanations (4) Our method is Fisher’s consistent even under either censoring model or survival model misspecification (but not both). This method has potential applications in multiple therapeutic areas. One direct impact for Diabetes business unit is that how we can leverage Lilly Diabetes’ broad treatment options to reduce or delay diabetes comorbidities such as CV event, diabetes related retinopathy, nephropathy, or neuropathy.

Statistical Method for the Analysis of Pooled Biomarker Data
Abby Sloan and *Molin Wang
Harvard T.H. Chan School of Public Health
stmow@channing.harvard.edu

For many health outcomes, it has become increasingly common to aggregate data from multiple studies to obtain increased sample sizes. The enhanced sample size of the pooled data allows investigators to perform subgroup analyses, evaluate the dose-response relationship over a broad range of exposures, and provide robust estimates of the biomarker-disease association. However, study-specific calibration processes must be incorporated in the statistical analyses to address between-study variability in the biomarker measurements. We introduce methods for evaluating the biomarker-disease relationship that validly account for the calibration process. We consider both internal and external calibration studies in the context of nested case-control studies. We then illustrate the utility of these estimators using simulations and an application to pooled data.

A zero-inflated logistic model for human microbiome data
Zhigang Li
Dartmouth College
zhigang.li@dartmouth.edu

Massive high dimensional human microbiome data is commonly seen in molecular epidemiology research and have substantially increased in complexity to address critical health concerns due to complex data structure. Analysis challenges arise from compositional, phylogenetically hierarchical, sparse and high dimensional structure of microbiome data. Compositional structure could in-
duce spurious relationships due to the linear dependence between compositional components. In addition, the hierarchical structure of microbiome data from the phylogenetic tree generates dependence at the hierarchical levels which poses a further modeling challenge. Furthermore, the sparsity of microbiome data due to excessive zero sequencing reads for microbial taxa remains an unresolved issue in the literature. Coupled with the high dimensional feature, microbiome data raises great challenging problems in the field of mediation data analysis. We develop a zero-inflated logistic normal model to address these issues. A simulation study will show the performance of the approach and a real study example will be included as well.

Session 123: Utilizing Real World Evidence in Regulatory Decision: Statistical Considerations and Beyond

The Use of Real-World Evidence for Making Regulatory Decisions
Greg Campbell
GCStat Consulting
patgregcampbell@verizon.net

In this era of “Big Data” there are many sources of possible data that could be leveraged in the regulatory environment. The effort to convert such data into evidence and then utilize such evidence in an efficient manner can be a challenge but the payoff can be enormous in terms of savings of time and resources. There are however inherent biases (especially selection biases) of observational data that need to be addressed as well as the confounders to be adjusted for due to different population characteristics. Also at issue is the generalizability of the studies. The potential uses of such real-world evidence in the premarket regulatory environment are: expansion of an indication, data for a control group using propensity score methodology in a prospective manner, an empirical prior of a Bayesian study, a registry based randomized trial and the setting of Objective Performance Criteria (OPC) or a Performance Goal (PG). Real-world evidence can also be useful in the postmarket in mandated surveillance studies (called 522 studies) and in condition-of-approval post market studies. The use of real-world evidence is being facilitated in the U.S. through a medical device National Evaluation System for health Technology (NEST) coordinated by the Medical Device Innovation Consortium (MDIC), a public-private partnership.

Incorporating Real World Evidence for Regulatory Decision Making with Propensity Score Methodology
Lilly Yue
U.S. FDA
lilly.yue@fda.hhs.gov

We are now at an amazing time for medical product development in drugs, biological products and medical devices. Dramatic recent advances in biological science, information technology and engineering as well as continuously accelerating trends toward globalization, make “big data” available in real world health care. However, data are not equal to knowledge or evidence. But, high-quality data can be transformed into scientific evidence through the application of proven analytical methods and techniques, such as propensity score methodology. This presentation will discuss opportunities and challenges in leveraging real world evidence using propensity score methods for regulatory decision making, with examples based on our regulatory review experience.

Roadmap from Real World Data to Real World Evidence: Reg-

ulatory Experience with Orthopedic Devices
Jianxiong Chu
Food and Drug Administration (FDA)
jianxiong.chu@fda.hhs.gov

Recently, we at FDA/CDRH have seen an increasing number of pre-market submissions relying on Real-World Data (RWD) collected from sources outside of traditional clinical trials (e.g., orthopedic device registries). In order for such RWD to be translated into Real World Evidence (RWE) for our regulatory decision making, there are many challenging issues remaining. For an example, often times, when a submission come in to CDRH, there is no assurance whether the real world data was collected and analyzed in a prospective way. In this talk, some basic principles and ideas for how to appropriately derive RWE from RWD will be shared based on our review experience using examples in the area of orthopedic devices.

Use of Real World Evidence in Cardiovascular Device Studies: A Statistical Reviewer’s Experience
♦ Heng Li and Vandana Mukhi
CDRH/Food and Drug Administration
heng.li@fda.hhs.gov

Real-world evidence, in particular data from registries, has served various purposes in informing regulatory decision-making. In the area of cardiovascular medical devices, registry data have been used directly and as controls in non-randomized comparative studies. In this presentation we will be mainly discussing the design of observational studies in the context of leveraging real-world evidence, i.e. what processes to follow in order to maintain study integrity and objectivity. We will provide details of the logistics in the implementation of those processes.

Session 124: Uncertainty, Effect Size and Bias in Statistical Evidence

P-values, bias and variability: a solution for replication power
♦ Laura C. Lazzeroni
Professor of Psychiatry and, by courtesy, of Medicine and Biomedical Data Science, Stanford University
lazzeroni@stanford.edu

Many statisticians have recently criticized the use of p-values. Despite their textbook definition as probabilities, p-values are random variables. This under-appreciated fact is one reason for many misunderstandings concerning their interpretation. Goodman was the first to discuss the relationship between observed p-values and predicted power in the replication setting under the alternative hypothesis. However, these replication-probability estimators are biased, with the worst bias often appearing under the null hypothesis. I will describe this problem and propose a solution, using a general-purpose, bias-correction procedure that I call the Ricochet. I will show how these predictions can improve the study design of genomic or other studies with multiple hypothesis tests and provide scientific researchers with realistic predictions of future study results.

Fair prediction with disparate impact: A study of bias in recidivism prediction instruments
Alexandra Chouldechova
Heinz College, Carnegie Mellon University
achould@cmu.edu

Recidivism prediction instruments (RPI’s) provide decision makers with an assessment of the likelihood that a criminal defendant will
reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses several fairness criteria that have recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when a recidivism prediction instrument fails to satisfy the criterion of error rate balance.

Beyond p-values: a phase II design with statistical significance and clinical relevance
Satrajit Roychoudhury
Novartis Pharmaceutical Corporation
satrajit.roychoudhury@novartis.com

Background
Well-designed Phase II trials must have acceptable error rates relative to a pre-specified success criterion, usually a statistically significant p-value. Such standard designs may not always suffice from a clinical perspective. Clinical relevance may call for more. For example, proof-of-concept in phase II often requires statistical significance and a sufficiently large effect estimate. We propose a dual-criterion design that complements statistical significance with clinical relevance. The aims of the paper are to describe the design and methodology of dual-criterion phase II trials and to illustrate their implementation in real life.

Methods
Clinical relevance requires the effect estimate to pass a clinically motivated threshold (the decision value). In contrast to standard designs, the required effect estimate is explicit whereas study power is implicit. The sample size for a dual-criterion needs careful considerations of the study’s operating characteristics (type-I error, power), which can be assessed analytically or with simulations.

Results
We have successfully applied dual-criterion designs in phase II. Frequentist and Bayesian phase II trials show various aspects of dual-criterion designs. These include decision criteria, sample size calculations, decisions under various data scenarios, and operating characteristics. We discuss implementations of dual-criterion designs in Oncology phase II trials with time-to-event and binary endpoints. The dual-criterion designs facilitate GO/NO-GO decisions due to their complementary statistical-clinical criterion.

Conclusion
To improve evidence-based decision-making, a formal quantitative framework is proposed. Dual-criterion designs offer an appealing statistical-clinical compromise. They may be preferable to standard designs if strong evidence against the null hypothesis alone does not suffice for an efficacy claim.

Session 125: Recent experience in accelerated approval in Oncology Immunotherapy from Industrial and FDA perspective

Accelerated Approval of Keytruda in Advanced Melanoma: The Pain and the Glory
• Nicole (Xiaoyun) Li and Cong Chen
Merck & Co.
chuche@iupui.edu

Anti-PD-L1 cancer therapy Keytruda received its first FDA approval in September 2014 in advanced melanoma, less than 4 years after the first human dosed with this drug. Even though the clinical development started more than 4 years later than its competitor, Keytruda was approved as first in class in the United States. In this presentation, we would like to summarize the study designs of the Keytruda melanoma program to support the accelerated approval and its conversion to full approval, share some experience on the interactions with regulatory agencies during review, and provide an outlook on the impact of the accelerated approval to the Keytruda program and the overall anti-PD1 landscape.

Recent experience in accelerated approval in Oncology from FDA perspective
• Vivian Yuan, Kun He and Rajeshwari Sridhara
FDA
vivian.yuan@fda.hhs.gov

The FDA instituted its Accelerated Approval Program to allow for earlier approval of drugs that treat serious conditions, and that fill an unmet medical need based on a surrogate endpoint. A surrogate endpoint is a marker, such as a laboratory measurement, radiographic image, physical sign or other measure that is thought to predict clinical benefit, but is not itself a measure of clinical benefit. The use of a surrogate endpoint can considerably shorten the time required prior to receiving FDA approval. The applicants are still required to conduct studies to confirm the anticipated clinical benefit. In this talk we will discuss statistical issues involved in this procedure and present some examples.

Session 126: Topics on Applied Statistics (II)

Assessing skeletal growth rate around pubertal growth spurt
• Cheng Hao Chu, Ying Zhang and Wanzhu Tu
Department of Biostatistics, IUPUI
chuche@iupui.edu

Human individuals acquire their adult body shapes through vigorous physical growth in the first two decades of life. Many of the somatic characteristics that define our physical appearance in adulthood take shape around the time of pubertal growth spurt (PGS), a time at which the velocity of height increase reaches its maximum. Little is known, however, about growth rates in other skeletal dimensions around the time of PGS, and whether these changes differ by subject characteristics. An analytical challenge to quantify growth rates before and after PGS is the lack of direct observation of the anchoring PGS event. In this research, we propose a linear model with interval-censored change point to assess the rates of skeletal changes around PGS. The model is nonparametric in the sense of not relying on parametric assumptions for the PGS timing and error distribution. A least-squares based method is used to estimate the model parameters, including the pre and post-PGS growth rates. We show that under mild regularity conditions, the estimators are consistent and asymptotically normal. Statistical inference ensues from the large sample theory. We conduct a simulation study to evaluate the operating characteristics of the proposed method. Analysis of growth data from an observational cohort shows that in comparison to girls, boys tend to have a more sustained skeletal growth after PGS, as evidenced by the greater post-PGS growth rates in the upper body and shoulder lengths, as well as in elbow, knee, and wrist diameters. The findings suggest that strong and sustained post-PGS skeletal growth contributes to the sexual dimorphism in human body.

Benefit and Risk Assessment in Ophthalmic Device Clinical
Studies
Chul Ahn
FDA
Chul.Ahn@FDA.HHS.GOV

Assessing the benefit-risk ratio of medical products in clinical studies can be a difficult task, because benefit and risk are two distinct variables. However, in some situations where they can be measured in the same unit, we may find an effective way of calculating a benefit-risk ratio. One such example is for ophthalmic devices intended to provide improvement in uncorrected (without glasses or contact lenses) near visual acuity, where uncorrected distance visual acuity may be used to assess risk, if subjects may be giving up distance vision for some gain in near vision. In this situation, one joint analysis of the benefit and risk of the device may be the assessment of how much distance vision the subjects give up (risk) for how much improvement in their near vision (benefit). In this paper, we propose a method for assessing this benefit and risk that involves calculation of a ratio, and the confidence intervals for our proposed ratio will be calculated based on Delta, Fieller, and Bayesian methods.

Method Comparison for Diagnostic Devices that Measure Different Parameters
Changhong Song
Food and Drug Administration
changhong.song@fda.hhs.gov

The objectives of many clinical studies are to demonstrate the equivalence of two diagnostic devices. If the two diagnostic devices measure the same parameter, an agreement analysis can be performed to evaluate whether there is systematic bias between the two devices. If the two devices measure similar or different parameters, the primary interest may be whether the two devices have equivalent linear trend with a clinical outcome. This study reviews some common statistical analysis methods that evaluate diagnostic devices that measure different parameters.

Some Considerations for the Trade-off Assessment of Diagnostic Errors when Comparing Diagnostic Test
Norberto Pantoja-Galicia and Gene Pennello
US Food and Drug Administration/CDRH
norberto.pantoja-galicia@fda.hhs.gov

This paper discusses the evaluation of the performance of a new diagnostic test in comparison with an already established test. We propose factors to be considered such as the benefits and risks from increased adherence to testing, the implicit or explicit trade-off between false positive and false negative test errors provided by the information from the Receiver Operating Characteristic (ROC) curve, as well as regulatory challenges and possible solutions.

Readmission Time Analysis for Psychiatry Patients Based on a Cox Intensity Process Model for Recurrence
Qing Li and Gideon Zamba
The University of Iowa
qing-li@uiowa.edu

Inpatient hospitalizations account for one-third of the annual healthcare costs in the USA. It is to be noted however that a large percentage of hospital readmissions can be avoided or prevented. At the University of Iowa Hospitals and Clinics, a nurse-led transitional care team (TCT) intervention is deployed in order to prevent unnecessary hospital readmissions. TCT is designed in a way to provide patients with disease self-management, medical education and clear instructions regarding discharge and hospital revisit. In this study we explore the effect of TCT intervention versus a Control, in a quasi-randomization type of analysis based on propensity score matching. In a Cox intensity process model adapted to recurrent events, we analyzed the inter-readmission times and explored the gap-time survival differences between TCT and Control.

Key Words: Recurrent Events, Cox Model

Rank Selection for Multilinear Principal Component Analysis
Ji-Ping Tu1, Su-Yun Huang2 and Dai-Ni Hsieh3
1Institute of Statistical Science, Academia Sinica
qiping@stat.sinica.edu.tw

Multilinear Principal Component Analysis (MPCA) has been a popular method to reduce the dimension of tensor data. One key challenge for further statistical applications of MPCA is to choose a proper tensor rank. Both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) need the likelihood function of the observation data that more strong model assumptions than the original model are required. In this paper, we propose a rank selection criterion for MPCA based on the minimum risk criterion and employ Stein’s unbiased risk estimate (SURE). We present SURE for the MPCA model in a neat formula composed of a residual sum for model fitting and a penalty on the model complexity expressed as the generalized degrees of freedom (GDF). We are able to match each term of the GDF to either the number of parameters used in the model or the complexity in separating the signal from the noise. Furthermore, SURE for the Principal Component Analysis model becomes a special case in our criterion. The simulation shows that, even for the data generated according to the assumptions of AIC and BIC, the selection method based on SURE reaches higher selection accuracies with much less computation loading, when the matrix size is large and the sample size is moderate.

Session 127: Topics on Clinical Statistics

Device Trials: Study Design Considerations
Rama Melkote
Janssen R&D
rmelkote@its.jnj.com

In designing clinical trials, a common method that is used to minimize bias and enhance validity of a statistical comparison of treatment groups is by random assignment of study treatment in a blinded manner; however, such blinding may not be feasible in device trials. In many device trials, in order to establish the clinical utility, the device is used as the experimental arm, which is then compared to a control arm where the device is not used. In these cases, a subject-level randomization may not be feasible as the use of the device in the experimental arm might influence the behavior of the clinician when making evaluations in the control arm. As a result, subject-level randomization might not show a statistical separation of the treatment arms. Additionally, maintaining the blind is also not feasible. Therefore, a traditional double-blind, parallel or cross-over study might not be the best design in these trials. The current presentation revolves around study designs that can be considered in designing a device study. Advantages, disadvantages and analysis approaches of each design will be discussed.

Semiparametric Model and Inference for Spontaneous Abortion Data with a Cured Proportion and Biased
Jin Piao1, Jing Ning2, Christina Chambers3 and Ronghui Xu4
1University of Texas Health Science Center
2The University of Texas MD Anderson Cancer Center
3University of California San Diego
jin.piao@uth.tmc.edu
Evaluating and understanding the risk and safety of using medications for autoimmune disease in a woman during her pregnancy will help both clinicians and pregnant women to make better treatment decisions. However, utilizing spontaneous abortion (SAB) data collected in observational studies of pregnancy to derive valid inference poses two major challenges. First, the data from the observational cohort are not random samples of the target population due to the sampling mechanism. Pregnant women with early SAB are more likely to be excluded from the cohort, and there may be substantial differences between the observed SAB time and those in the target population. Second, the observed data are heterogeneous and contain a ‘cured’ proportion. In this article, we consider semiparametric models to simultaneously estimate the probability of being cured and the distribution of time to SAB for the uncured subgroup. To derive the maximum likelihood estimators, we appropriately adjust the sampling bias in the likelihood function and develop an expectation-maximization algorithm to overcome the computational challenge.

We apply the empirical process theory to prove the consistency and asymptotic normality of the estimators. We examine the finite sample performance of the proposed estimators in simulation studies and illustrate the proposed method through an application to SAB data from pregnant women.

**Protocol deviations in medical device pivotal clinical studies**

♦ Zhiheng Xu and Meijuan Li
US Food and Drug Administration
zhiheng.xu@fda.hhs.gov

The golden standard in clinical study practice is to follow what it is pre-specified in the study protocol. However, there are situations where deviations from the protocol are occurred. For example, the investigator may develop new inclusion/exclusion criteria during the conduct of trial. When enrolling patient, slow accrual could lead to unplanned patient expansion and low observe event rate could result in unplanned enrichment. Those protocol deviations can expose the study to various biases of unknown size and direction, and biases can adversely impact the level of evidence provided by the study and the ability to rely on the data as valid. In this talk, we will discuss the issues of protocol deviations and provide the approaches in minimizing bias and adjusting for patient enrichment in medical device pivotal clinical studies.

**Optimal design theory in late-onset toxicity problem**

Tian Tian
AbbVie
tiantian234823@gmail.com

One of the practical problems encountered in early-phase oncology clinical studies is the late-onset toxicity. Here we propose a design called the OD-CRM where optimal design theory is incorporated into the standard structure of the CRM. We build in the ITTE weighting mechanism into the design where it assigns each toxicity response to a weight depending on the patient’s enrollment time and the observed data. We also offer a general dose-finding algorithm based on the OWEA (optimal weight exchange algorithm), Yang, Biedermann, and Tang, 2013, to explore the performance of the OD-CRM under a broader clinical trial setup.

**Unraveling the Secret of Biologic Drugs’ Usage in the U.S.**

♦ Yunyue Zhang, Giorgio Quer, Steve R. Steinhubl, Eric J. Topol and Nathan E. Wineinger
Scripps Translational Science Institute
parlitz@scripps.edu

Biologic pharmaceuticals are the top-selling drugs with a substantially high per-patient cost. Globally, biologic sales have increased from 11

We reviewed blinded administrative claims data from January 2013 through June 2016 from anonymous insurers conducting business throughout the United States, including over 35 million unique monthly individuals with pharmacy coverage. In this database, 1,577,149 Humira prescription claims were paid, representing 152,466 unique individuals. The number of claims increased annually from 27,803 per month in 2013 to 40,785 per month in 2016. At the same time, the average per-claim cost increased from 3,047 in 2013 to 4,829 in 2016. Results from data extracted from health insurance claims and other high-cost biologics are ongoing. These results will produce a national, large-scale inspection of variation biologic prescriptions across temporal, geographic, and demographic levels, and address questions related to biologic costs and clinical utility. We expect this will inform future discussions on biologic use in medicine.

**The Lead Time Distribution in the National Lung Screening Trial Study**

♦ Ruiqi Liu¹, Adriana P??rez² and Dongfeng Wu³
³University of Louisville
r0liu009@louisville.edu
²UT Health Science Center at Houston

Objectives: Lung cancer screening with low-dose computed tomography is recommended for heavy smokers. The effectiveness of lung cancer screening exams regarding the survival benefits becomes a major concern since the lead time is unobserved. In this study, the estimates for the projected lead time was presented by using the National Lung Screening Trial (NLST) computed tomography (CT) arm data.

Methods: A developed probability model for the lead time distribution was applied where the length of a person’s lifetime was considered as a random variable rather than a fixed value. The lifetime distribution was obtained based on the actuarial life table available from the United States Social Security Administration (2016). Then the lead time distribution for participants in the CT arm was projected using the NLST data. Four initial screening ages (55, 60, 65 and 70) were examined with four screening intervals (12, 18, 24 and 30 months). For each scenario, lead time distribution for both screening detected cases and interval cases were estimated.

Results: Simulation results show that the probability of no-early detection increases monotonically when the screening interval increases for both genders. For example, a male and female heavy smoker with initial screening age at 60 has 11.65% decrease when the screening interval increases but it seems stable across different initial age groups.

Conclusions: The projected lead time for participants in a lung cancer screening program with low-dose CT scans was obtained. This result is hoped to provide a statistical evidence for policy makers on the evaluation of effectiveness for lung cancer screening with CT scans.

**Session 128: Statistical Analysis in High-Dimensional Data Analysis with Applications**

**An ensemble method for RNA-Seq differential analysis**

♦ Dongmei Li, Martin Zand, Thomas Fogg, Zidian Xie and Timothy Dye
University of Rochester
donmeli11@urmc.rochester.edu

Abstracts
Confident Inference for SNP Effects on Treatment Efficacy

Han Chen and Xihong Lin

Harvard T.H. Chan School of Public Health

With the advance in next-generation sequencing technology, massive genetic and genomic data have been produced. These data are often sparse, as most genetic variants in the human genome are rare mutations with very low minor allele frequencies. Statistical methods for testing genetic association with rare variants have been proposed and widely applied to unrelated samples. These methods are also known as gene-based or variant-set tests, since rare variants are often grouped by genes or genomic regions in the analysis. The burden test and the sequence kernel association test (SKAT) are two types of widely applied rare variant tests. Here we propose and implement the burden test, SKAT and a hybrid test for structured and correlated samples. All tests share the same null generalized linear mixed model, which only needs to be fitted once in a whole-genome analysis. We show in simulation studies that the proposed tests control correct type I error rates in the presence of population stratification and cryptic relatedness, in both single-cohort studies and meta-analysis. We compare the power of these tests in various scenarios and illustrate how they can be used to test a broad class of different scientific hypotheses in large-scale sequencing studies. We also apply the methods to a real data example.

Association tests for sparse genetic data in structured and correlated samples

Han Chen and Xihong Lin

Harvard T.H. Chan School of Public Health

RNA-Seq differential analysis is widely used in biological and biomedical research to identify genes associated with interested traits or diseases. Numerous statistical methods have been proposed for RNA-Seq differential analysis and implemented in open source statistical analysis software such as R/Bioconductor. Unfortunately, the false discovery rate (FDR) of all current methods are much higher than the nominal level for small sample size RNA-Seq experiments, making further validation of selected significant genes be prone to high probability of false discoveries. To reduce the false discovery rate in gene differential analysis for RNA-Seq experiments with small sample size, we propose an ensemble method using four current methods (DESeq2, EBSeq, SAMSeq, NOISeq) with equal weight on their FDR adjusted P-values. Simulation studies showed our proposed ensemble method had much less FDR than current popular methods and could control the FDR within nominal level even when sample size was very small such as three in each group. The application of our ensemble method to real RNA-Seq experimental data significantly reduced the probability of false discoveries as shown in our real data applications, which could help increase the confidence of biological and biomedical researchers when they conduct verifications of selected significant genes.

Detecting Genetic Subgroups of Patients in a Tailored Drug Development Process

Ying Ding and Jason Hsu

1University of Pittsburgh

2Ohio State University

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 cost effective treatment of patient subgroups in a tailored drug development process.

Multiple Imputation for Missing Data in Analgesic Clinical Trials using Pattern Mixture Models

Xueya Cai, Michael McDermott, Jennifer Gewandter, Hua He and Robert Dworkin

University of Rochester

Tulane University

Missing data in longitudinal clinical trials are unavoidable. In clinical trials of pain treatments, the percentage of participants who withdraw early can be as high as 50% depending on the medication, dosage, and duration of the study. Major reasons for early withdrawal in these studies include perceived lack of efficacy and adverse events. An analysis based on the missing at random (MAR) assumption may not be useful, especially for withdrawals due to adverse events. Current strategies for accommodating missing data in analgesic clinical trials include last observation carried forward (LOCF), baseline observation carried forward (BOCF), “responder” analyses, and methods that assume MAR, including linear mixed effects models, mixed model repeated measures (MMRM) analyses, and multiple imputation. None of the existing applications considers different missingness mechanisms for different reasons for early withdrawal, which may be more realistic in analgesic clinical trials. In this study, we implemented a multiple imputation strategy for missing data based on pattern-mixture modeling (PMM), as described by Carpenter et al., and applied it to data from clinical trials of pregabalin for treatment of several different chronic pain conditions. We used a “jump-to-reference” strategy to impute missing data for trial participants who withdrew due to adverse events and an MAR assumption for participants who withdrew for other reasons. Estimated treatment effects are presented and compared with those from other existing strategies for accommodating missing data.

Session 129: Biomarker Integration in Cancer Clinical Trials

Statistical Issues in Biomarker Clinical Trials

Sumithra Mandrekar

Mayo Clinic

As cancer has become increasingly understood on the molecular level, newer “targeted” drugs that inhibit specific cancer cell growth and survival mechanisms have increased the need for new clinical trial designs, wherein pertinent questions on the relationship between patient biomarkers and response treatment can be addressed. Increasingly common are trials tailored to detect enhanced efficacy in a patient subpopulation, e.g. patients with a known biomarker value or whose tumors harbor a specific genetic mutation. Diagnostics validation is a very crucial and essential part of these clinical studies to verify the validity and generalizability of the assay methodology and biomarker tests used for molecularly targeted
therapies. This talk will discuss the systematic evaluation for the proper clinical validation of tailored tests and treatments, as well as provide an overview of the development of new design strategies, both for early phase and definitive trials, using examples of real clinical clinical trials.

An adaptive design for the identification of the optimal dose using joint modelling of biomarker measurements and toxicity over all treatment cycles in phase I/II clinical trials of molecularly targeted agents in oncology

Maria-Athina Alzernakou1 and Xavier Paoletti2
1 CESP OncoStat, INSERM, University of Paris-Saclay
2 Institut Gustave Roussy, CESP OncoStat

Background: Conventional dose-finding approaches in oncology of phase I clinical trials aim to identify the maximum tolerated dose (MTD), based on the toxicity events observed during the first treatment cycle. Even though this is relevant for cytotoxic agents, this may not be the case for molecularly targeted agents, usually administered in chronic schedules. Lately, continuous biomarkers are used more and more to monitor activity. However, activity does not necessarily increase monotonically with dose. Therefore, both toxicity and activity should be considered for the identification of the optimal dose (OD). The OD will be defined as the lowest dose that achieves high activity, while satisfying certain toxicity requirements; this may not be the MTD.

Methods: We present an adaptive dose-finding method, based on a joint modelling technique of a longitudinal outcome for continuous biomarker activity measurements and a probit time to first severe toxicity, with shared random time slope. This method allows for exact likelihood inference (Barrett J. et al., 2015), an important property, in the context of small sample sizes, typical of phase I/II trials. We combined the aforementioned technique with the time to event continual reassessment method that used data collected over all treatment cycles. The dose allocated to a new patient was the lowest dose in the same range as the most active dose, under the constraint of being below or equal to the MTD. For the evaluation of the method we ran a set of simulations covering a wide range of scenarios. Dose and time were included in both models, with the addition of dose-time interactions for the longitudinal outcome. We considered up to 6 dose levels, 6 treatment cycles, a maximum of 18 visits for the biomarker measurements and 60 patients.

Results: In our scenarios the correct identification of the OD ranged between 91% and 100%, with the exception of one scenario, in which the correct OD selection was 72%. Regarding the safety of the design, there was only one scenario where the algorithm would suggest as OD the dose right above the MTD, and that was in less than 2% of the cases. In other words, it is very rare that highly toxic doses will be selected to proceed to the next phases. We have also investigated scenarios that analyze data of both biomarker measurements and time to event toxicity that come from different models; this may not necessarily increase monotonically with dose. Therefore, both toxicity and activity should be considered for the identification of the optimal dose (OD). The OD will be defined as the lowest dose that achieves high activity, while satisfying certain toxicity requirements; this may not be the MTD.

Methods: We present an adaptive dose-finding method, based on a joint modelling technique of a longitudinal outcome for continuous biomarker activity measurements and a probit time to first severe toxicity, with shared random time slope. This method allows for exact likelihood inference (Barrett J. et al., 2015), an important property, in the context of small sample sizes, typical of phase I/II trials. We combined the aforementioned technique with the time to event continual reassessment method that used data collected over all treatment cycles. The dose allocated to a new patient was the lowest dose in the same range as the most active dose, under the constraint of being below or equal to the MTD. For the evaluation of the method we ran a set of simulations covering a wide range of scenarios. Dose and time were included in both models, with the addition of dose-time interactions for the longitudinal outcome. We considered up to 6 dose levels, 6 treatment cycles, a maximum of 18 visits for the biomarker measurements and 60 patients.

Results: In our scenarios the correct identification of the OD ranged between 91% and 100%, with the exception of one scenario, in which the correct OD selection was 72%. Regarding the safety of the design, there was only one scenario where the algorithm would suggest as OD the dose right above the MTD, and that was in less than 2% of the cases. In other words, it is very rare that highly toxic doses will be selected to proceed to the next phases. We have also investigated scenarios that analyze data of both biomarker measurements and time to event toxicity that come from different models; this may not necessarily increase monotonically with dose. Therefore, both toxicity and activity should be considered for the identification of the optimal dose (OD). The OD will be defined as the lowest dose that achieves high activity, while satisfying certain toxicity requirements; this may not be the MTD.

Discussion: Overall, our design performs very well especially when the optimal dose is on the tail of the dose range. It has the ability to deal with missing at random measures, which is an important feature in this very advanced population, expected to have rapid deterioration due to the progressive disease.
Objective: The statistical approach recommended in previous FDA/CDER guidances for generic transdermal delivery systems products (TDS) had a low power for “well-adhering” TDS, impacting availability of generic TDS to the public. In 2016, CDER published a draft guidance to resolve this issue.

Methods: Mathematical proof and simulation were used to evaluate the cause of the low power of the previous statistical non-inferiority (NI) approach. A new NI hypothesis was proposed and a NI margin was determined based on simulation and real data analyses. Power was compared among different approaches.

Results: Regarding the cause for the low power of the historical statistical approach for well-adhering TDS, one common consideration was that the non-normality of the adhesion data (i.e., the high skewness and the discrete feature of ordinal scores) violated the normality assumption of the linear mixed model used to evaluate NI, and therefore, might cause the low passing rate for well-adhering TDS products. Extensive statistical research revealed that non-normality of the adhesion data was not actually the true cause of the problem. Rather, it was determined that the direction of the adhesion scale (a smaller score indicates better adhesion: 0 for perfect adhesion, 4 for complete detachment) coupled with the use of a ratio of the mean (ROM) scores (Test/RLD) for the NI statistical test were the main causes for the low power (passing rate) of the historical statistical approach for well-adhering TDS products. A new statistical hypothesis was recommended in the draft guidance on adhesion by replacing the traditional ratio-of-means NI test with a difference-of-means (DOM) NI test, still based upon mean adhesion scores. DOM NI test is robust in power to the direction of adhesion scores (whereas ROM NI test is highly sensitive to it), and can dramatically improve the power for “well-adhering” TDS products. The NI margin of 0.15 for the difference of means was determined to be appropriate based upon collaborative research by clinicians, statisticians, and other scientists.

Conclusion: The currently recommended statistical approach in the new guidance corrects the low power, i.e., the low passing rate of the historical statistical approach recommended by the previous product-specific guidances for well-adhering TDS while retaining the targeted type 1 (false positive) error rate under 0.05. It is also consistent with previously passing TDS.

Statistical Considerations of Adhesion Data Analysis

Yu-Ting Weng, Caroline Strasinger and Meiyu Shen
FDA
Yu-Ting.Weng@fda.hhs.gov

The draft guidance of assessing adhesion with transdermal delivery systems and topical patches for abbreviated new drug applications (ANDAs) was issued by FDA in 2016. In the guidance, the primary endpoint is the mean adhesion score based on a 5-point numerical scaled score that corresponds to a specified range of adhered surface area of the transdermal delivery system. For primary endpoint analysis, the non-inferiority test is used to demonstrate adequate adhesion performance of the test product compared to the reference product. However, currently no guidance is available for evaluating the adhesion performance of new drugs in new drug applications (NDA). Thus, in this study for the primary endpoint, we discuss the advantage and disadvantage of using the ordinal-scaled score and propose to use the continuous scaled score in NDA reviews. For primary endpoint analysis, a test-product-only approach based on the superiority test is proposed to evaluate the adhesion performance of new drugs in NDA reviews and how to determine an appropriate margin is provided for the discussion.
We will present results in sparse linear regression on two convex regularized estimators, the Lasso and the recently introduced Slope estimator, in the high-dimensional setting where the number of covariates is larger than the number of observations. The estimation and prediction performance of these estimators will be presented, as well as a comparative study of the assumptions on the design matrix.

Nonparametric Methods in Business Analytics
Haipeng Shen
University of Hong Kong
shanhaipeng@gmail.com

Business analytics refers to data-driven decision making in business. Nonparametric methods are playing an increasing role in business analytics, fueled by increasing data complexity and dimensionality. I shall showcase the power of nonparametric methods using real business applications.

Factor Models for Matrix-Valued High-Dimensional Time Series
Dong Wang, Xiaolu Liu and Rong Chen
1Princeton University
2San Diego State University
3Rutgers University
dongwang@princeton.edu

to data-driven decision making in business. Nonparametric methods are playing an increasing role in business analytics, fueled by increasing data complexity and dimensionality. I shall showcase the power of nonparametric methods using real business applications.

Session 132: High dimensional inference and its application to change-point analysis

High-dimensional covariance matrix estimation
Yilei Wu, Yingli Qin and Mu Zhu
University of Waterloo
yingli.qin@uwaterloo.ca

We study high-dimensional covariance matrix estimation under the assumption of the low-rank and diagonal matrix decomposition. The covariance matrix estimator is decomposed into a low-rank matrix L and a diagonal matrix D, and the rank of L is either fixed to be small or suppressed by a penalty function. Under moderate conditions on the population covariance matrix and the penalty function, our estimator enjoys some consistency results. An algorithm which iteratively updates L and D is applied to solve for the estimator. Some simulations and real data analysis are presented to show the performance with finite sample size.

Simultaneous Inference for Multiple Change points
Ning Hao
University of Arizona
nhao@math.arizona.edu

As a classical topic, change-point analysis is experiencing a renaissance due to recent applications to big and noisy sequence data from genomics and econometrics among other fields. In this talk, we briefly review recent breakthroughs on this topic and discuss a local approach to multiple change-point detection, which is computationally efficient and theoretically verified. Moreover, we illustrate a framework to conduct simultaneous inference and control the false discovery rate.

Test for temporal homogeneity of high-dimensional means with application to fMRI studies
Jun Li and Ping-Shou Zhong
1Kent State University
2Michigan State University
jli49@kent.edu

Motivated by the region of interest analysis in the fMRI studies, we consider the problem of testing temporal homogeneity of p-dimensional population mean vectors from the repeated measurements of n subjects over T times. To cope with the challenges brought by fMRI data, we propose a test statistic that takes into account not only the large p, large T and small n situation, but also the complex temporospatial dependence of the fMRI data. The asymptotic distribution of the proposed test statistic is established under mild conditions. When the null hypothesis of temporal homogeneity is rejected, we further propose a binary segmentation method shown to be consistent for multiple change-points identification. An application to fMRI data is provided to demonstrate the performance of the proposed methods.

Sparse Multivariate Statistics with Discrete Optimization
Rahul Mazumder
Massachusetts Institute of Technology

Several statistical estimation tasks arising in modern multivariate statistics are naturally posed as discrete optimization problems. While continuous convex optimization methods have played a highly influential role in these tasks, the role of modern discrete optimization methods, namely, integer programming has been relatively less explored, despite the tremendous advances in the field over the past 10-15 years. In this talk, I will describe how techniques in modern optimization: mixed integer optimization, first order methods in nonlinear optimization provides a systematic algorithmic lens to address some key problems in sparse multivariate statistics. I will illustrate this new approach with examples in variable selection in regression, robust statistical regression, function estimation and factor analysis.

Session 133: Leadership and Career Development

Statistical Leadership and Career Development in Pharmaceutical Industry
Ivan Chan
AbbVie, Inc.

The overarching mission of the pharmaceutical industry is to develop medical products that can improve human health. Scientific discoveries and innovations in research and development are very critical in the success of developing new medical products such as drugs, vaccines, and medical devices. Working closely with scientists and clinical researchers, statisticians play a key role in solving real problems in the drug development process, from basic discovery to clinical trials and post marketing surveillance. In this presentation, we will share some thoughts on the statistical career and the
leadership role that statistician can play in pharmaceutical development. In addition, we will highlight the key leadership qualities that will help statisticians advance their career in the collaborative research environment.

**Career Development for Biostatisticians in Academia**

*Haitao Chu*

University of Minnesota  
chux0051@umn.edu

With the burgeoning need for (big) data analyses, the world is full of exciting professional opportunities for biostatisticians. In this talk, we will discuss some popular academic career paths for those with training or interests in biostatistics. Which every research career path is different; there are some common grounds on how to maximize your chances of making it as a permanent academic. While “more publications” is certainly critical, it is not the whole story. Securing external funding (particularly in a soft money environment), teaching and student advising, research reputation, personal and professional reputation can all influence your chances. As last, I will share my own experience and career path from a medical student in China with little mathematical training to a tenured full professor in Biostatistics at the University of Minnesota Twin Cities, with an emphasis on effective collaborations with other biomedical investigators.

**My Leadership Experience in Government**

*Yi Tsong*

CDER, FDA  
yi.tsong@fda.hhs.gov

In my personal opinion, leadership requires the following qualities: A clear vision; Never stop learning; Not afraid of changing; Always acknowledge people; Accept mistakes made; Never stop caring; Passion to your work. In government, statisticians may play many roles of leadership. These roles include Team leader, expert reviewer or scientist, Division director; Office director. Unofficially, these roles also include member and leader of research project, statistical topic working group, FDA Guidance committee. In this presentation, I will share with all participants my 30 years personal professional development as a statistician and leader in FDA.

**Session 134: Recent Advances in Analytical Methods for Cancer Genomics**

**Probabilistic models and statistical methods in cancer etiology and evolution**

*Cristian Tomasetti*

Johns Hopkins University  
crtomasetti@jhu.edu

Cancers are caused by mutations that may be inherited (H), induced by environmental factors (E), or the result of random mistakes made during normal DNA replication (R). A fundamental problem in cancer etiology is the estimation of the proportion of mutations attributable to each one of these three factors. In fact, these proportions have never been estimated. I will present recent results in the mathematical, probabilistic modeling of cancer evolution that allow, when combined with statistical methods and epidemiological information, the above estimation. The results suggest that R mutations play a major role in cancer causation (1). Using a completely orthogonal approach, based solely on cancer genome sequencing and a statistical method, we provide novel mutational signatures in cancer and again estimate the role of R in cancer.

Work in collaboration with Lu Li, Bert Vogelstein, andBahman Af-sari (Johns Hopkins University).


**Quantifying tumor evolution via spatial computational modeling and Approximate Bayesian Computing**

*Christina Curtis*

Stanford University School of Medicine  
cncurtis@stanford.edu

Cancer results from the acquisition of somatic alterations in an evolutionary process that typically occurs over many years, much of which is occult. Understanding the evolutionary dynamics that are operative at different stages of progression in individual tumors might inform the earlier detection, diagnosis, and treatment of cancer. Although these processes cannot be directly observed, the resultant spatiotemporal patterns of genetic variation amongst tumor cells encode their evolutionary histories. For example, we recently described a “Big Bang” model of human colorectal tumor growth, whereby after transformation, the neoplasm grows predominantly as a single expansion in the absence of stringent selection and where the timing of a mutation is the fundamental determinant of its frequency in the final tumor. By analyzing multi-region genomic data within a spatial agent-based tumor growth model and Approximate Bayesian Computing framework, we demonstrated the early origin of intra-tumor heterogeneity and delineated the dynamics of tumor growth in a patient-specific manner. The Big Bang model is compatible with effectively neutral evolution and suggests that not all tumors exhibit stringent selection after transformation, thereby challenging the de facto clonal expansion model. These findings emphasize the need for the systematic evaluation of different modes of evolution across different tumor types and methods to infer the role of natural selection in established human tumors. To address this need, we developed an extensible framework to simulate spatial tumor growth and evaluate evidence for different modes of tumor evolution. Application of this approach to multi-region sequencing data from diverse tumor types reveals different evolutionary modes and times with implications for how human tumors progress and ultimately how they may be more effectively treated.

**MEGSA: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations**

♦ Xing Hua, Paula Hyland, Jing Huang, Bin Zhu, Neil Caporaso, Maria Landi and Jianxin Shi

National Cancer Institute  
huax@mail.nih.gov

The central challenge in tumor sequencing studies is to identify driver genes and pathways, investigate their functional relationships and nominate drug targets. The efficiency of these analyses, particularly for infrequently mutated genes, is compromised when patients carry different combinations of driver mutations. Mutual exclusivity analysis helps address these challenges. To identify mutually exclusive gene sets (MEGS), we developed a powerful and flexible analytic framework based on a likelihood ratio test and a model selection procedure. Extensive simulations demonstrated that our method outperformed existing methods for both statistical power and the capability of identifying the exact MEGS, particularly for highly imbalanced MEGS. Our method can be used for de novo discovery, pathway-guided searches or for expanding established small MEGS. We applied our method to the whole exome sequencing data.
for fourteen cancer types from The Cancer Genome Atlas (TCGA). We identified multiple previously unreported non-pairwise MEGS in multiple cancer types. For acute myeloid leukemia, we identified a novel MEGS with five genes (FLT3, IDH2, NRAS, KIT and TP53) and a MEGS (NPM1, TP53 and RUX1) whose mutation status was strongly associated with survival ($P = 6.7 \times 10^{-4}$). For breast cancer, we identified a significant MEGS consisting of TP53 and four infrequently mutated genes (ARID1A, AKT1, MED23 and TBL1XR1), providing support for their role as cancer drivers.

Keywords: Mutual exclusivity, oncogenic pathways, driver genes, tumor sequencing

Session 135: Recent developments in statistical machine learning

Multilayer tensor factorization with applications to recommender systems

* Xuan Bi¹, Annie Qu² and Xiaotong Shen³

¹Yale University
²University of Illinois at Urbana-Champaign
³University of Minnesota Twin Cities

xuan.bi@yale.edu

Recommender systems have been widely adopted by electronic commerce and entertainment industries for individualized prediction and recommendation, which benefit consumers and improve business intelligence. In this article, we propose an innovative method for tensor recommender systems. The proposed method utilizes the structure of a tensor response to integrate information from multiple modes, and creates an additional layer of nested latent factors to accommodate between-subjects dependency. One major advantage is that the proposed method is able to address the "cold-start" issue in the absence of information from new customers, new products or new contexts. Specifically, it provides more effective recommendations through sub-group information. In theory, the proposed method achieves an optimal convergence rate under the $L_2$-loss function and a fast convergence rate in more general settings. To achieve scalable computation, we develop a new algorithm for the proposed method, which incorporates the maximum block improvement strategy into the blockwise coordinate descent. Finally, the proposed method is applied in simulations and IRi marketing data with 116 million observations of product sales. Numerical studies demonstrate that the proposed method significantly outperforms existing competitors in the literature.

A new SVD approach to optimal topic estimation

* Tracy Ke and Minzhe Wang

University of Chicago

zke@galton.uchicago.edu

In the probabilistic topic models, the quantity of interest "a low-rank matrix consisting of topic vectors" is hidden in the text corpus matrix, masked by noise, and Singular Value Decomposition (SVD) is a potentially useful tool for learning such a low-rank matrix. However, the connection between this low-rank matrix and the singular vectors of the text corpus matrix are usually complicated and hard to spell out, so how to use SVD for learning topic models faces challenges.

We overcome the challenge by revealing a surprising insight: there is a low-dimensional simplex structure which can be viewed as a bridge between the low-rank matrix of interest and the SVD of the text corpus matrix, and which allows us to conveniently reconstruct the former using the latter. Such an insight motivates a new SVD-based approach to learning topic models.

For asymptotic analysis, we show that under the popular probabilistic topic model (Hofmann, 1999), the convergence rate of the $\ell_1$-error of our method matches that of the minimax lower bound, up to a multi-logarithmic term. In showing these results, we have derived new element-wise bounds on the singular vectors and several large-deviation bounds for weakly dependent multinomial data. Our results on the convergence rate and asymptotical minimaxity are new. We have applied our method to two data sets, Associated Process (AP) and Statistics Literature Abstract (SLA), with encouraging results. In particular, there is a clear simplex structure associated with the SVD of the data matrices, which largely validates our discovery.

On the connections between algorithmic regularization and penalization for GLM's

* Qian Qian, Vincent Vu and Yanzhong Zhu

The Ohio State University

qian.216@osu.edu

It is known that the iterates of the gradient descent algorithm for least squares linear regression can be viewed as generalized ridge regression estimators. This establishes a formal link between algorithmic regularization and penalization that can be used for understanding early stopping. We show that an analogous link can be made between penalized estimation of generalized linear models and the natural gradient descent algorithm. The corresponding penalties involve Bregman divergences between the model parameter and its initialization, encouraging shrinkage toward the initialization. Such a connection is useful in two ways. First, it allows a new perspective on the algorithmic regularization for a larger class of models. Second, for the purpose of model selection, algorithmic regularization often tends to be computationally more efficient than penalization.

Session 136: Recent advances in the analysis of complex data

Shape Constrained Tensor Factorizations

* Eric Chi¹, Bethany Lusch² and Nathan Kutz²

¹North Carolina State University
²University of Washington

eric.chi@ncsu.edu

We consider N-way data arrays and low-rank tensor factorizations where the time mode is coded as a sparse linear combination of temporal elements from an over-complete library. Our method, Shape Constrained Tensor Decomposition (SCTD) is based upon the CANDECOMP/PARAFAC (CP) decomposition which produces r-rank approximations of data tensors via outer products of vectors in each dimension of the data. By constraining the vector in the temporal dimension to known analytic forms which are selected from a large set of candidate functions, more readily interpretable decompositions are achieved and analytic time dependencies discovered. The SCTD method circumvents traditional flattening techniques where an N-way array is reshaped into a matrix in order to perform a singular value decomposition. A clear advantage of the SCTD algorithm is its ability to extract transient and intermittent phenomena which is often difficult for SVD-based methods. We motivate the SCTD method using several intuitively appealing results before applying it on a number of high-dimensional, real-world data sets in order to illustrate the efficiency of the algorithm in extracting interpretable spatio-temporal modes. With the
rise of data-driven discovery methods, the decomposition proposed provides a viable technique for analyzing multitudes of data in a more comprehensible fashion.

Matrix Linear Discriminant Analysis
Wei Hu1, Weining Shen1, Hua Zhou2 and Dehan Kong3
1University of California, Irvine
2University of California, Los Angeles
3University of Toronto

We propose a novel linear discriminant analysis approach for the classification of high-dimensional matrix-valued data that commonly arises from imaging studies. Motivated by the equivalence of the conventional linear discriminant analysis and the ordinary least squares, we consider an efficient nuclear norm penalized regression that encourages a low-rank structure. Theoretical properties including a non-asymptotic risk bound and a rank consistency result are established. Simulation studies and an application to electroencephalography data show the superior performance of the proposed method over the existing approaches.

A Note on Inverse Regressions When Responses are Missing at Random
Yuexiao Dong, Cheng Yong Tang and Qi Xia
Temple University
tuf14573@temple.edu

Inverse regressions constitute a class of sufficient dimension reduction methods targeting at estimating the central space by regression-type approaches implemented inversely on the predictors and the responses. The most representative approach in this family is the seminal sliced inverse regression (SIR) approach proposed by Li (1991). In this study, we first show that missing responses generally affect the validity of the inverse regressions under the scheme of the so-called missing at random, in the sense that the resulting estimations for the central space can be biased if observations with missing response are ignored. We then propose a simple and effective adjustment for missing response that guarantees the validity of the inverse regressions. The proposed method share the essence and simplicity of the classic SIR. Furthermore, a marginal coordinate test is introduced for the proposed estimator. We demonstrate the performance of the proposed inverse regressions for dealing with missing responses by theoretical and numerical analyses.

Nonparametric estimation of multivariate mixture
Chaowen Zheng and Yichao Wu
North Carolina State University
czheng6@ncsu.edu

A multivariate mixture model is determined by three elements: the number of components, the mixing proportions and the component distributions. Assuming that we are given the number of components and that each mixture component has independent marginal distributions, we propose a non-parametric algorithm to estimate the component distributions in a multivariate mixture model. The basic idea is that we convert the estimation of density functions as a problem of estimating the coordinates of density functions under a good basis of functions. Specifically, we construct a set of basis functions by some conditional density functions and try to recover the coordinates of component distributions under this basis. In the simulation study, we compare our algorithm with other existing non-parametric methods of estimating component distributions in mixture models under the assumption of conditionally independent marginals.

Session 137: Statistical Inference for High-dimensional Linear Regression and Covariance Structure

Structured Volatility Matrix Estimation with Accurate Co-volatility Estimators based on Non-synchronous
Jinqing Fan, Dongyu Kim and Kun Lu
Princeton University

Recently several large volatility matrix estimation procedures have been developed based on factor-based It processes which have an integrated volatility matrix consisting of low-rank and sparse matrices. Their performances depend on the accuracy of input volatility matrix estimators. In practice, when estimating co-volatilities, we run into the non-synchronization problem which even more severe for illiquid assets. Due to this, it is hard to expect that the co-volatility estimators for illiquid assets represent their co-volatilities well.

In this paper, we investigate how to estimate the large integrated volatility matrix without co-volatility estimators for illiquid assets. Specifically, we pretend that the co-volatilities for illiquid assets are missing, and we estimate the low-rank matrix using a matrix completion scheme with the structured missing pattern. Then using the adaptive thresholding scheme, we estimate the sparse matrix. We also investigate the asymptotic properties of the proposed estimation procedure. To check the finite sample performance, we conduct extensive simulation study and apply the proposed estimation procedure to the real high-frequency financial data.

Selection of Effective Scores for Treatment Effect Estimation
Lei Wang, Ying Zhang, Menggang Yu and Jun Shao
University of Wisconsin-Madison
lewang.stat@gmail.com

Estimation of average and quantile treatment effects is crucial in causal inference for evaluation of treatments or interventions in biomedical, economic, and social studies. Under the assumption of treatment and potential outcomes are independent conditional on all covariates, valid treatment effect estimators can be obtained using nonparametric inverse propensity weighting and/or regression, which are popular because no model on propensity or regression is imposed. To obtain valid and efficient treatment effect estimators, typically the set of all covariates can be replaced by lower dimensional sets containing linear combinations of covariates, which are called effective scores. We propose to construct an effective score separately for each treatment and show that the resulting asymptotic variance of treatment effect estimator reaches a lower bound that is smaller than those based on other effective scores. Since the effective scores have to be estimated, for example, using nonparametric sufficient dimension reduction, we derive theoretical results on when the efficiency of treatment effect estimation is affected by estimating effective scores. We find that, except for some special cases, the efficiency of treatment effect estimation is affected even though the sufficient dimension reduction is consistent in the rate of the square root of the sample size. Our theory is complemented by some simulation results and an illustration is also made using data from the University of Wisconsin Health Accountable Care Organization.

Structured Correlation Detection with Application to Colocalization Analysis in Microscopic Imaging
Shulei Wang1, Jianqing Fan2, Ginger Pocock3 and Ming Yuan1
1University of Wisconsin-Madison
2Princeton University
shulei@stat.wisc.edu
Motivated by the problem of colocalization analysis in fluorescence microscopic imaging, we study in this paper structured detection of correlated regions between two random processes observed on a common domain. We argue that although intuitive, direct use of the maximum log-likelihood statistic suffers from potential bias and substantially reduced power, and introduce a simple size-based normalization to overcome this problem. We show that scanning with the proposed size-corrected likelihood ratio statistics leads to optimal correlation detection over a large collection of structured correlation detection problems.

CHIME: Clustering of High-Dimensional Gaussian Mixtures with EM Algorithm and Its Optimality

Tony Cai, Jing Ma and Linjun Zhang
University of Pennsylvania
Linjunz@wharton.upenn.edu

Unsupervised learning is an important problem in statistics and machine learning with a wide range of applications. In this paper, we consider clustering of high-dimensional Gaussian mixtures and propose a procedure, called CHIME, that is based on the EM algorithm and a direct estimation method for the sparse discriminant vector. Both theoretical and numerical properties of CHIME are studied. We first obtain the rates of convergence for the estimation and clustering errors and then provide matching minimax lower bounds. The results together establish the optimality of CHIME as well as the proposed estimators of the discriminant vector and other parameters. Simulation studies show that CHIME outperforms the existing methods under a variety of settings. The proposed CHIME procedure is also illustrated in an analysis of a glioblastoma gene expression data set and shown to have superior performance. Clustering of Gaussian mixtures in the conventional low-dimensional setting is also considered. The technical tools developed for the high-dimensional setting are used to establish the optimality of the clustering procedure based on the classical EM algorithm.

Session 138: Recent advances in statistical genomics

Practical Issues of meta-analysis of genetic associations using sequence data

Yu Jiang
Penn State College of Medicine
dxl46@psu.edu

There is great interest to understand the impact of rare variants in human diseases using large sequence datasets. Meta-analysis has been a valuable tool to enlarge sample sizes and empower genetic association studies. As genetic studies move from using genotyping arrays to sequence data, new practical issues arise where further method development is in great need. In this talk, we focus on improved methods for meta-analysis in the presence of multi-allelic variants and missing summary association statistics from participating studies. Existing meta-analysis methods either replace missing summary statistics with zero or discard studies with missing data. These naive approaches can bias genetic effect estimates and lead to seriously inflated type-I or II errors for conditional analysis. To address this widespread issue, we developed a method to borrow strength across participating studies and consistently estimate the joint effects for candidate and known variants. Based on the consistent estimator, we propose a pseudo-score statistic for conditional analysis of single variant and gene-level associations. The new method produces well-calibrated type I errors and is substantially more powerful than existing approaches in the presence of missing summary statistics. We illustrated the effectiveness of the proposed methods using extensive simulation studies and applied them in a large-scale meta-analysis of nicotine addiction phenotypes. We expect these new developments to be beneficial for the next phase large scale meta-analysis of sequence data.

A graphical model to prioritizing GWAS results guided by biomedical literature mining

Dongjian Chung1, Hang Kim2, Zhenning Yu3, Andrew Lawson1 and Hongyu Zhao3

1Medical University of South Carolina
2University of Cincinnati
3Yale University
chungd@musc.edu

Genome-wide association studies (GWAS) have successfully identified genetic variants associated with various diseases. However, complex diseases are often affected by a large number of genetic variants with small to moderate effect sizes, and the identification of these risk variants remains a challenging problem. Recently, integration of biomedical big data has been considered as a powerful approach to address these challenges. In this presentation, I will discuss our novel statistical approach for the joint analysis of multiple GWAS data sets using a graphical modeling approach guided by a text mining of biomedical literature. Our study was motivated by the accumulating evidence suggesting that different complex diseases share common genetic basis, i.e., pleiotropy. In addition, integration of biomedical literature mining results allows us to utilize prior biological knowledge to improve the estimation of pleiotropic architecture. This approach does not only increase statistical power to identify risk variants, but also provides a parsimonious representation of genetic relationship among phenotypes, within a unified framework. I will illustrate the proposed method with simulation studies and application to real GWAS datasets.

Structured subcomposition selection in regression and its application to microbiome data analysis

Tao Wang
Shanghai Jiao Tong University
neowangtao@sjtu.edu.cn

Compositional data arise naturally in many practical problems and the analysis of such data presents many statistical challenges, especially in high dimensions. In this talk, we consider the problem of subcomposition selection in regression with compositional covariates, where the relationships among the covariates can be represented by a tree with leaf nodes corresponding to covariates. Assuming that the tree structure is available as prior knowledge, we adopt a symmetric version of the linear log contrast model, and propose a tree-guided regularization method for this structured subcomposition selection. Our method is based on a novel penalty function that incorporates the tree structure information node-by-node, encouraging the selection of subcompositions at subtree lev-
els. We show that this optimization problem can be formulated as a generalized lasso problem, the solution of which can be computed efficiently using existing algorithms. An application to a human gut microbiome study and simulations are presented to compare the performance of the proposed method with an $l_1$ regularization method where the tree structure information is not utilized.

**AC-PCA: simultaneous dimension reduction and adjustment for confounding variation**

Zhixiang Lin, Can Yang, Hongyu Zhao and Wing Hung Wong

Stanford University

Hong Kong Baptist University

Yale University

Dimension reduction methods are commonly applied to high-throughput biological datasets. However, the results can be hindered by confounding factors, either biological or technical in origin. In this study, we extend Principal Component Analysis to propose AC-PCA for simultaneous dimension reduction and Adjustment for Confounding variation. We show that AC-PCA can adjust for (a) variations across individual donors present in a human brain exon array dataset, and (b) variations of different species in a model organism ENCODE RNA-Seq dataset. Our approach is able to recover the anatomical structure of neocortical regions, and to capture the shared variation among species during embryonic development.

For gene selection purposes, we extend AC-PCA with sparsity constraints, and propose and implement an efficient algorithm. The methods developed in this paper can also be applied to more general settings. The R package and MATLAB source code are available on https://github.com/linzx06/AC-PCA.

**Session 139: Analysis of High-Dimensional Genomics Data**

Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach

Dungang Liu and Heping Zhang

University of Cincinnati Lindner College of Business

Yale University School of Public Health

Ordinal outcomes are common in scientific research and everyday practice, and we often rely on regression models to make inference. A long-standing problem with such regression analyses is the lack of effective diagnostic tools for validating model assumptions. The difficulty arises from the fact that an ordinal variable has discrete values that are labeled with, but not, numerical values. The values merely represent ordered categories. In this paper, we propose a surrogate approach to defining residuals for an ordinal outcome Y. The idea is to define a continuous variable S as a “surrogate” of Y and then obtain residuals based on S. For the general class of cumulative link regression models, we study the residual’s theoretical and graphical properties. We show that the residual has null properties similar to those of the common residuals for continuous outcomes. Our numerical studies demonstrate that the residual has power to detect misspecification with respect to 1) mean structures; 2) link functions; 3) heteroscedasticity; 4) proportionality; and 5) mixed populations. The proposed residual also enables us to develop numeric measures for goodness-of-fit using classical distance notions. Our results suggest that compared to a previously defined residual, our residual can reveal deeper insights into model diagnostics. We stress that this work focuses on residual analysis, rather than hypothesis testing. The latter has limited utility as it only provides a single p-value, whereas our residual can reveal what components of the model are misspecified and advise how to make improvements.

**Non-parametric Empirical Bayes method for sparse, noisy signals**

Junhui Cai, Ya’acov Ritov and Linda Zhao

University of Pennsylvania

University of Michigan

lzhao@wharton.upenn.edu

Genomic data often comes in a form which is noisy and sparse. It is challenging to recover the truth due to its complex structure and high dimensionality.

We propose to use Bayesian nonparametric schemes to tackle the problem. The method adapts especially well to varying degrees of sparsity. It not only performs well to recover the signals, but also provides credible intervals. We also propose a method to control FDR in the case of multiple testing.

**False Discovery Rate Control for High-Dimensional Networks of Conditional Quantile Associations**

Jichun Xie and Ruoshua Li

Duke University

The University of Texas at Houston

jichun.xie@duke.edu

Motivated by the gene co-expression pattern analysis, we propose a novel squac statistic to infer quantile associations conditioning on covariates. It features enhanced flexibility in handling variables with both arbitrary distributions and complex association patterns conditioning on covariates. We first derive its asymptotic null distribution, and then develop a multiple testing procedure based on squac to simultaneously test the independence between one pair of variables conditioning on covariates for all $p(p+1)/2$ pairs. Here, p is the length of the outcomes and could exceed the sample size. The testing procedure does not require resampling or perturbation, and thus is computationally efficient. We prove by theory and numerical experiments that the squac testing method asymptotically controls the false discovery rate. It outperforms all alternative methods when the complex association patterns exist. Applied to a gastric cancer data, the squac method estimated the gene co-expression networks of early and late stage patients. It identified more changes in the networks which are associated with cancer survivals. We extend our method to the case that both the length of the outcomes and the length of covariates exceed the sample size, and show that the asymptotic theory still holds.

**Session 140: Recent Development of Statistical Learning Methods for Precision Medicine**

Subgroup Identification from Multiple Individual Patient Level Data Sets

Chensheng Kuang, Menggang Yu, Qi Tang, Alan Hartford and Weinig Robieson

University of Wisconsin-Madison

Sanofi

ckuang@wisc.edu

We consider subgroup identification based on multiple studies with individual patient data. Compared with exploratory subgroup analysis from a single study, our effort may alleviate concerns that include insufficient power and poor generalizability. We adopt a weighted
contrast classification framework to the meta-analysis. Instead of modeling outcomes, this new framework focuses on treatment classification while using outcomes as weights. We use penalties that exploit similarities among the studies but also allow study-to-study heterogeneity. We evaluate our method by comparing with three outcome based modeling approaches in simulation and a real data analysis.

**VARIABLE SELECTION FOR ESTIMATING THE OPTIMAL TREATMENT REGIMES**

*Baqun Zhang* 1 and *Min Zhang* 2

1 Shanghai University of Finance and Economics, China  
2 University of Michigan  

mzhangst@umich.edu

Most of existing methods for optimal treatment regimes, with few exceptions, focus on estimation and are not designed for variable selection with the objective of optimizing treatment decisions. In clinical trials and observational studies, often numerous baseline variables are collected and variable selection is essential for deriving reliable optimal treatment regimes. Although many variable selection methods exist, they mostly focus on selecting variables that are important for prediction (predictive variables) instead of variables that have a qualitative interaction with treatment (prescriptive variables) and hence are important for making treatment decisions. We propose a variable selection method within a general classification framework to select prescriptive variables and estimate the optimal treatment regime simultaneously. In this framework, a optimal treatment regime is equivalently defined as the one that minimizes a weighted misclassification error rate and the proposed method forward sequentially select prescriptive variables by minimizing this weighted misclassification error. A main advantage of this method is that it specifically targets selection of prescriptive variables and in the meantime is able to exploit predictive variables to improve performance. The method can be applied to both single- and multiple-decision point setting. The performance of the proposed method is evaluated by simulation studies and application to a clinical trial.

**Inference for Two-Stage Dynamic Treatment Regimes in the Presence of Drop-Out**

*Abdus Wahed* 1 and *Andrew Topp* 2

1 Dept. of Biostatistics, University of Pittsburgh  
2 Abbvie  
wahed@pitt.edu

Statistical Inference for dynamic treatment regimes from observational studies or randomized trials are generally done via inverse-probability-of-treatment-weighting or Robin’s g-Computation. G-computation is a straightforward approach that fits a regression model of the outcome as a function of the treatment along with observed patient characteristics and then combines the appropriate treatment subgroup means using total probability formula. Inverse probability weighting instead estimates a patient’s probability of receiving treatment and weights the outcomes by the inverse of this probability. In this talk, we discuss these techniques and then extend them to estimate the effects of dynamic treatment regimes in the presence of drop-out that are missing at random. We propose a new inverse probability of treatment weighted estimator that accounts for such drop-outs, and derive it’s variance. We compare variations of this method with the existing methods via simulation. We utilize the new class of estimators in the analysis of the STAR*D trial to estimate the mean outcome of patients on different regimes for the treatment of non-psychotic major depressive disorder.

**Statistical Learning Methods to Improve Risk Prediction in Organ Transplantation**

*Lihui Zhao*  
Northwestern University  
lihui.zhao@northwestern.edu

Liver transplantation (LT) is currently the only definitive treatment for patients with end stage liver disease. A major challenge in LT is the growing disparity between organ supply and waitlist demand. Currently, there are over 10,000 new patients with ESLD added to LT waiting list each year, but only 6,000 LT are performed each year in the US. Optimizing donor-recipient matching can improve liver utilization and transplant outcomes. In this talk, we will present several statistical learning methods to improve risk prediction models for patient and graft survival outcomes of individual donor-recipient matching. The application to the data from the Organ Procurement and Transplant Network will be illustrated.

### Session 141: New developments and Challenges in Biomarker Discovery

**Protein Biomarker Discovery from ELISA Cytokines Multiplex Assays**

*Veavi Chang*  
Eli Lilly and Company  
veavi0106@gmail.com

Cytokines are a large group of proteins, peptides or glycoproteins that are secreted by specific cells of immune system. Profiling expression of multiple cytokines is essential to unraveling the mechanisms and effects of many disease processes. A growing demand of cytokines biomarker is applied in clinical trials for early disease detection and disease diagnosis. Currently the cytokine data is analyzed for each cytokine separately and often ignore the dependence structure from the same subject. We propose a statistical framework to visualize, summarize, and analyze not only for each cytokine separately but also integrate multiple cytokine profiles together. Ultimately, the proposed method combines cytokine profiles with other phenotypes primary endpoints for new biological findings.

**Detection for the Onset of Alzheimer’s Disease Using Joint Modeling**

*Tian Wang and Jimin Ding*  
Washington University in St. Louis  
tianwang@wustl.edu

Alzheimer’s disease (AD) is the most common type of degenerative dementia, and the increasing prevalence of AD in this century intensifies the need for greater research efforts. Given the irreversible nature of the disease, it is especially important to detect AD at the earlier stage to treat the disease effectively. In this talk, I will use a two-step joint model to investigate the longitudinal cognitive changes and facilitate the diagnosis of early AD. In the first step, the annually assessed cognitive item responses are modeled by the generalized linear mixed effects models (GLMM) to characterize the features of longitudinal impairment, while adjusting the demographic covariates. In the second step, the fitted random effects, which represent the extracted individual feature, are then incorporated in the survival models to predict the early stage of AD. Here, the onset of diagnosis of mild cognitive impairment (MCI) is defined as the event time. The hazard rates are modeled based on the demographic covariates, baseline cognitive scores, and the individual random effects from the GLMM.

**Scalable Analysis for Mass Spectrometry Imaging Experiments**

*Kylie Bemis and Olga Vitek*
Enhancement of the Adaptive Signature Design for Learning and Confirming in a Single Pivotal Trial

Gu Mi
Eli Lilly and Company

Because of the complexity of cancer biology, often the target pathway is not well understood at the time that Phase III trials are initiated. Friedlin and Simon [1] proposed a two-stage trial design for identifying a subgroup of interest in a learn stage, based on one or more baseline biomarkers, and then subsequently confirming it in a confirmation stage. In this article, we discuss some practical aspects of this type of design and describe an enhancement to this approach that can be built into the study randomization to increase the robustness of the evaluation. Furthermore we show via simulation studies how the proportion of patients allocated to the learn stage versus the confirm stage impacts the power and provide recommendations.

A non-randomized procedure for discrete multiple testing based on randomized tests

XIAOYU DAI, Nan Lin, Daofeng Li and Ting Wang
Washington University in St. Louis

In the detection of differentially methylated regions (DMRs) with the analysis of next-generation sequencing technology, massive discrete data are generated from short read counts and lead to multiple discrete testing. However, most existing multiple testing procedures controlling false discovery rate (FDR) assume that test statistics are continuous and become conservative when applied to the DMR detection. To increase the power of DMR detection, in this article, we propose a novel multiple testing procedure for better FDR control on discrete tests. Our procedure makes decisions based on the marginal critical function (MCF) of randomized tests, utilizing which enables achieving a powerful and non-randomized multiple testing procedure. We further prove upper bounds of the positive FDR (pFDR) and the positive false non-discovery rate (pFNR) for our procedure, and show that the pFDR is controlled at the nominal level under mild conditions. We also prove that the set of detections made by our method contains every detection made by a naive application of the q-value method. We further demonstrate the improvement of our method over other existing methods by simulations and a real DMR detection analysis using a whole-genome bisulfite sequencing dataset.

Session 142: Topics on Applied Statistics (III)

Optimal Number of Tissue Specimens Required for Diagnosis of Prosthetic Joint Infection.

Jay Mandrekar
Mayo Clinic

We recently demonstrated improved sensitivity of prosthetic joint infection (PJI) diagnosis using an automated blood culture bottle system for periprosthetic tissue culture. Current study builds on the prior research by examining the optimal number of periprosthetic tissue specimens required for accurate PJI diagnosis. Current guidelines recommend five to six, which is impractical. We applied Bayesian latent class modeling techniques for estimating diagnostic test properties of conventional culture techniques (aerobic and anaerobic agars and thioglycolate broth) compared to inoculation into blood culture bottles. Conventional, frequentist receiver operating characteristic curve analysis was conducted as a sensitivity analysis. Results of this study show that the greatest accuracy of PJI diagnosis is obtained when three periprosthetic tissue specimens are obtained and inoculated into blood culture bottles or four periprosthetic tissue specimens are obtained and cultured using standard plate and broth cultures. Increasing the number of specimens to five or more, per current recommendations, does not improve accuracy of PJI diagnosis.

Higher moments modified VaR estimators and their applications in portfolio

Shu-Hui Yu
National University of Kaohsiung

Recently, some financial crises, for example, financial tsunami and European debt crisis, push governments around the world to make some new regulations to control the risk better. How to estimate VaR more accurately receives increasing attentions in researchers and practitioners. Accordingly, there are some higher moments modified VaR estimations to reflect the asymmetric and heavy-tailed characteristics of financial data and to accommodate the new regulations. Therefore, considering the time series dependence, clustering, and asymmetric characteristics in data, we assume data come from a GJR model in this talk. And we will derive the higher moments modified VaR estimations by maximizing expected utility function, orthogonal estimating function, and percentile expansion. Also, we will apply the results to find the underlying assets in the optimal portfolio. The main purpose of this talk is to provide some new higher moments modified VaR estimations. Also, we will apply these results to optimal portfolio problems.

Permutation Test Based on a Linear Mixed Model for Differential Alternative Splicing Analysis

Huining Kang
University of New Mexico

Changes in gene expression can correlate with poor disease characteristics and outcomes in two ways: through changes in relative transcript levels or through alternative RNA splicing leading to changes in relative abundance of individual transcript isoforms. There is an increasing evidence that the transcript isoform expression profiles provide more informative cancer signatures than standard gene expression profiles. RNA-sequencing technologies provide powerful tools for transcriptome analysis, reconstruction and quantification and offer an unprecedented opportunity to discover
novel genes, transcripts and splice variants underlying complex diseases. However, statistical methods for detection of differentially expressed isoforms are still in their infancy and need development, especially in the context of more advanced experimental designs. One of the challenges is how to appropriately taken into account the correlations among the isoforms of one gene and that among multiple genes. We recently developed a linear mixed effects model for analyzing the complex alternative RNA splicing regulation patterns which naturally incorporates the important features of biological correlation patterns between isoforms of the same gene. In this presentation we further propose a permutation test for estimating the false discovery rates (FDR) which also takes into account the correlations among different genes. Based on an RNA-sequencing data set from a study of adenoid cystic carcinoma (ACC), we applied our approach to identify genes that have differentially expressed isoforms, and compared the results to those obtained from our original approach that uses the Benjamini-Hochberg’s FDR procedure.

Finding observed Fisher information matrix for EM based inference on misrepresentation risk in insurance

Rexford Akakpo1, Michelle Xia1 and Alan Polansky1

Misrepresentation (false statements) by policy applicants may increase the risk borne of insurance companies. Misrepresentation is unidirectional, meaning policy applicants give false statements in only one direction that is in their favor. This gives heterogeneity in modeling claim outcomes given certain risk factors. A natural way to estimate these is using conditional distribution of the observed that has a form similar to that of finite mixture models. We propose a misrepresentation model with equal variance, use the Expectation Maximization (EM) algorithm and missing information principle to find observed Fisher information for the model. We derive the algorithm, and the observed Fisher information matrix and implemented the algorithm in R.

Keywords: Misrepresentation, unidirectional, Expectation Maximization, Fisher information

Automatic Forecasting of Electricity Demand With a Computationally Efficient Semi-Parametric Model

Jun Liu

In this paper we develop a semi-parametric approach to model non-linear relationship in time series data. The usefulness of this approach is illustrated on a hourly electricity demand data set. Polynomial splines are used to model the effect of temperature on hourly electricity demand for different times of the day and types of the day. An ARIMA model is used to model the serial correlation in the data. An algorithm is developed to automatically select the models, and the models are estimated through backfitting. Forecasting performance is evaluated using post-sample forecasting and comparative results are presented.

Update on progress of ASA Biopharm Safety Monitoring Working Group

Susan Duke

To better promote public health and protect patient safety, the ASA Biopharm Safety Monitoring Working Group is developing a systematic approach for safety evaluation of pharmaceutical products. Our focus is pre-marketing safety monitoring. This talk is an update on the group’s progress. The presentation will touch on deliverables currently being developed, including a) a manuscript describing the key safety reporting guidance documents from ICH, US, EU, CIOMS, b) a cross-disciplinary update of the Program Safety Analysis Plan, and c) methodology recommendations, some using Bayesian techniques and others using interactive graphics.

The presentation will be of interest to those who wish to achieve clinical relevance, consistency, clarity, efficiencies, quantitative rigor and transparency in how our industry addresses drug safety during the drug development lifecycle. It’s part of the value proposition: Proactive safety monitoring for comprehensive benefit-risk evaluation. Goals of the presentation are to make biostatisticians aware of this WG’s effort and take feedback.

Session 143: New developments and statistical issues in evidence synthesis

Hierarchical Models for Combining N-of-1 Trials

Christopher Schmid

Brown University

N-of-1 trials are single-patient multiple-crossover studies for determining the relative effectiveness of treatments for an individual participant. A series of N-of-1 trials assessing the same scientific question may be combined to make inferences about the average efficacy of the treatment as well as to borrow strength across the series to make improved inferences about individuals. Series that include more than two treatments may enable a network model that can simultaneously estimate and compare the different treatments. Such models are complex because each trial contributes data in the form of a time series with changing treatments. The data are therefore both highly correlated and potentially contaminated by carryover. We will use data from a series of 100 N-of-1 trials in an ongoing study assessing different treatments for chronic pain to illustrate different models that may be used to represent such data.

Multiple dosage levels in network meta-analysis using the potential outcomes approach: A comparison

Mireille Schnitzer1 and Michele Bally2

Université de Montréal

Network meta-analysis is used to compare the effectiveness of more than two treatments using the results of previous studies. However, it is often noted that characteristics of the assigned treatments, such as dosage levels of medications, vary between studies. Usage of the potential outcomes framework provides a theoretical approach for incorporating multiple versions of treatment into the definition of the effect of interest in a network meta-analysis. Our results show that under several additional assumptions (including no interference between studies and study arms) consistent estimation of the causal effect of interest is possible when adjusting for all common causes of the treatment and treatment version in addition to all common causes of the treatment and the mean outcome in each study arm. We demonstrate the usage of G-Computation to adjust for study-level covariates to estimate the marginal effect of interest in a small-sample simulation study. This method is then employed to evaluate the relative effectiveness of four biological agents (and placebo) for the treatment of moderate to severe plaque psoriasis in a network meta-analysis.

The impact of excluding trials from network meta-analyses : an
Network meta-analysis (NMA) expands the scope of a conventional pairwise meta-analysis to simultaneously compare multiple treatments, which has an inherent appeal for clinicians, patients, and policy decision makers. Two recent reports have shown that the impact of excluding a treatment on NMAs can be substantial. However, no one has assessed the impact of excluding a trial from NMAs, which is important because many NMAs selectively include trials in the analysis. This article empirically examines the impact of trial exclusion using both the arm-based (AB) and contrast-based (CB) approaches, by reanalyzing 20 published NMAs involving 725 randomized controlled trials and 449,325 patients. For the population-averaged absolute risk estimates using the AB approach, the average fold changes across all networks ranged from 1.004 (with standard deviation 0.004) to 1.072 (with standard deviation 0.184); while the maximal fold changes ranged from 1.032 to 2.349. In 12 out of 20 NMAs, a 1.20-fold or larger change is observed in at least one of the population-averaged absolute risk estimates. In addition, while excluding a trial can substantially change the estimated relative effects (e.g., log odds ratios), there is no systematic difference in terms of changes between the two approaches. Changes in treatment rankings are observed in 7 networks and changes in inconsistency are observed in 3 networks. We do not observe correlations between changes in treatment effects, treatment rankings and inconsistency. Finally, we recommend rigorous inclusion and exclusion criteria, logical study selection process, and reasonable network geometry to ensure robustness and generalizability of the results of NMAs.

A Bayesian meta-analysis method for estimating risk difference of rare events
Yuanyuan Tang
Saint Luke’s Health System
ytutang@saint-lukes.org

Bayesian meta-analysis has been more frequently utilized for synthesizing safety and efficacy information to support landmark decision-making due to its flexibility of incorporating prior information and availability of computing software. However, when the outcome is binary and the events are rare, where event counts can be zero, conventional meta-analysis methods including Bayesian methods may not work well. Several methods have been proposed to tackle this issue but the prior knowledge of event rate was not utilized to increase precision of risk difference estimates. To better estimate risk differences, we propose a new Bayesian method, Beta prior Binomial model for Risk Differences (B-BIRD), which takes into account the prior information of rare events. B-BIRD is illustrated using a real data set of 48 clinical trials about a type 2 diabetes drug. In simulation studies, it performs well in low event rate settings.

Session 144: Recent Advances in Quantile Regression

A Flexible Dependence Measure for Semi-Competing Risks Data: An Application of Quantile Regression
Jing Yang and Limin Peng
Emory University
lpeng@emory.edu

Semi-competing risks data are often encountered in chronic disease follow-ups that record both nonterminal events (e.g., disease landmark events) and terminal events (e.g., death). Studying the relationship between the nonterminal event and the terminal event can provide insightful information on disease progression. In this talk, I will introduce a new sensible dependence measure tailored to addressing such an interest via a novel application of quantile regression. We define the new dependence measure and propose its nonparametric estimators within the framework of quantile regression. We further develop a regression method to adjust for covariates for the new dependence measure by utilizing its connection with quantile regression. We establish the asymptotic properties of the proposed estimators and develop inferences accordingly. Simulation studies suggest good finite-sample performance of the proposed methods. Our proposals are illustrated via applications to Denmark diabetes registry data.

Testing for Marginal Effects in Quantile Regression
Huixia Wang1, Ian McKeague2 and Min Qian2
1George Washington University
2Columbia University
judywang@gwu.edu

We develop a new marginal testing procedure to detect the presence of significant predictors associated with the quantiles of a scalar response. The idea is to fit marginal quantile regression models based on each predictor separately, then select the predictor that minimizes the empirical quantile loss function and use the corresponding slope estimate as a test statistic. A resampling method is devised to calibrate this test statistic, which has non-regular limiting behavior due to the variable selection. Asymptotic validity of the procedure is established in a general quantile regression setting in which the marginal quantile regression models can be misspecified. Even though a fixed dimension is assumed to derive the asymptotic results, the proposed test is applicable and computationally feasible for large-dimensional predictors. The method is more flexible than existing marginal screening test methods based on mean regression, and has the added advantage of being robust against outliers. The approach is illustrated using an application to an HIV drug resistance dataset.

Central Quantile Subspace
Eliana Christou
University of North Carolina at Charlotte
echrisl5@uncc.edu

Dimension reduction is a useful technique when working with high-dimensional predictors, where straightforward graphical analysis is not possible. Existing dimension reduction techniques focus on the conditional distribution of the response given the covariates, where specific interest focuses on statistical functionals of the distribution, such as the conditional mean, conditional variance, or conditional quantile. We introduce a new method for inferring about the conditional quantile of the response given the covariates and we introduce the notion of the Central Quantile Subspace (CQS). The purpose of this presentation is three fold. First, we focus on cases where the conditional quantile depends on the predictors X only through a single linear combination β′X and show that we can estimate β consistently up to a multiplicative scalar, even though the estimate might be based on a misspecified link function. Second, we extend our result to conditional quantiles that depend on the predictors X through a d-dimensional linear combination B′X, where B is a p × d matrix, d > 1, and propose an iterative procedure to produce more vectors in the CQS, which are shown to be root n consistent. Third and last, we extend our proposed methodology by considering any statistical functional of the conditional distribution and estimate the
fewest linear combinations of X that contain all the information on just that functional.

Session 145: Recent Developments in Statistical Methods in Healthcare

Precision Care of Stroke Patients: Trajectory Analysis of Dynamic Blood Pressure
Hui Zhi
University of Hong Kong

Stroke is a “chronic disease with acute events” that causes long-term healthcare issues. It is one of the leading causes of death and disability, for example, 1 in China and 5 in North America. Precision care of stroke patients has become a critical worldwide healthcare challenge. We use trajectory analysis of dynamic blood pressure data to suggest a more powerful approach to associate with medical outcome of interests, such as recurrence of stroke within one year after the initial stroke episode.

Tree based weighted learning for estimating individualized treatment rules with censored data
Yifan Cui1, Ruoqing Zhu2 and Michael Kosorok3
1University of North Carolina at Chapel Hill
2University of Illinois at Urbana-Champaign

Estimating individualized treatment rules is a central task for personalized medicine. Zhao et al. (2012) and Zhang et al. (2012) proposed outcome weighted learning to estimate individualized treatment rules directly through maximizing the expected clinical response. In this paper, we extend outcome weighted learning to right censored survival data without requiring either inverse probability of censoring weighting or semiparametric modeling of the censoring and failure times as done in Zhao et al. (2015b). To accomplish this, we take advantage of the tree based approach proposed in Zhu & Kosorok (2012) to nonparametrically impute the survival time in two different ways. The first approach replaces the reward of each individual by the expected survival time, while the second method imputes the expected failure time conditional on the observed censoring time. We establish consistency and convergence rates for both estimators. In simulation studies, our estimators demonstrate superior performance compared to existing methods. We also illustrate the proposed method on a phase III clinical trial of non-small cell lung cancer.

On Reject and Refine Options in Multicategory Classification
Chong Zhang1, Wenbo Wang2 and Xingye Qiao3
1Seattle WA
2Binghamton University

In many real applications of statistical learning, a decision made from misclassification can be too costly to afford; in this case, a reject option, which defers the decision until further investigation is conducted, is often preferred. In recent years, there has been much development for binary classification with a reject option. Yet, little progress has been made for the multicategory case. In this article, we propose margin-based multicategory classification methods with a reject option. In addition, and more importantly, we introduce a new and unique refine option for the multicategory problem, where the class of an observation is predicted to be from a set of class labels, whose cardinality is not necessarily one. The main advantage of both options lies in their capacity of identifying error-prone observations. Moreover, the refine option can provide more constructive information for classification by effectively ruling out implausible classes. Efficient implementations have been developed for the proposed methods. On the theoretical side, we offer a novel statistical learning theory and show a fast convergence rate of the excess L-risk of our methods with emphasis on diverging dimensionality and number of classes. The results can be further improved under a low noise assumption and be generalized to the excess O(d-1) risk. Finite-sample upper bounds for the reject and reject/refine rates are also provided. A set of comprehensive simulation and real data studies has shown the usefulness of the new learning tools compared to regular multicategory classifiers.

Predicting Hospital Admissions from Emergency Departments at Triage
Han Ye1, Zhankun Sun2 and Haipeng Shen3
1University of Illinois at Urbana Champaign
2City University of Hong Kong
3University of Hong Kong

We develop models that use patient information collected during triage at emergency departments to predict the inpatient admission decisions, and test how the information derived from free-text triage notes by text mining can increase the prediction power.

Session 146: Recent Advances on Neuroimaging Data Analysis

Optimal experimental designs for fMRI when the model matrix is uncertain
Ming-Hung Kao and Lin Zhou
Arizona State University

This study concerns optimal designs for functional magnetic resonance imaging (fMRI) experiments when the model matrix of the statistical model depends on both the selected stimulus sequence (fMRI design), and the subject’s uncertain feedback (e.g. answer) to each mental stimulus (e.g. question) presented to her/him. While practically important, this design issue is challenging. This mainly is because that the information matrix cannot be fully determined at the design stage, making it difficult to evaluate the quality of the selected designs. To tackle this challenging issue, we propose an easy-to-use optimality criterion for evaluating the quality of designs, and an efficient approach for obtaining designs optimizing this criterion. We show that our approach can significantly outperform a previously proposed method.

Simultaneous Confidence Bands for Mean and Variance Functions based on Deterministic Design
Li Cai1, Rong Liu2, Suojin Wang3 and Lijian Yang4
1Soochow University
2University of Toledo
3Texas A&M University
4Tsinghua University

Asymptotically correct simultaneous confidence bands (SCBs) are proposed for the mean and variance functions of nonparametric regression model based on deterministic designs. The variance estimation is as efficient as up to order \( n^{-1/2} \) as an infeasible estimator if the mean function were known. Simulation experiments provide
strong evidence that corroborates the asymptotic theory. The proposed methodology is illustrated via an application to two human event-related potentials (ERP) real data sets (visual spatial attention ERP data and metaphor semantic integration ERP data).

Spatially Varying Coefficient Models

Jingru Ma¹, Guannan Wang² and Lily Wang¹
¹ Iowa State University
² College of William & Mary

Spatially varying coefficient models have been a popular tool to explore the spatial non-stationarity of a regression relationship for spatial data. In this paper, we study the estimation of spatially varying coefficient models for data distributed over complex domains. We use bivariate splines over triangulations to represent the coefficient functions. The estimators of the coefficient functions are consistent, and rates of convergence of the proposed estimators are established. A penalized bivariate spline estimation method with more flexible choice of triangulations is also proposed. In addition, we propose hypothesis tests to examine if the coefficient function is really varying over space. The proposed method is much more computationally efficient than the well-known geographically weighted regression technique and thus usable for analyzing massive datasets. The performance of the estimators and the proposed tests are evaluated by a few simulation examples and a real data analysis.

Functional data analysis for dynamic PET data

R. Todd Ogden, Yuxuan Chen and Jeff Goldsmith
Columbia University, Department of Biostatistics, New York, NY
todd@columbia.edu

One major goal of dynamic positron emission tomography (PET) imaging, is the estimation of the spatial distribution of specific molecules throughout the brain. Common analysis strategies involve applying parametric models that require fairly strong assumptions, reducing information for each subject and each voxel/region into a single scalar-valued summary, and modeling each subject and each voxel/region sequentially. We will describe extensions of the analysis in three different directions: a nonparametric approach to the modeling of the observed PET data; a functional data analytic (FDA) approach to modeling the impulse response function; and the ability to consider observed PET data from multiple subjects in a single (“function-on-scalar”) regression model. We demonstrate the application of this approach and compare the results with those derived from standard parametric approaches.

Session 147: Statistical learning for complex data

High-dimensional Linear Regression for Dependent Data with Applications to Nowcasting

Yuefeng Han and Ruey Tsay
University of Chicago
ruey.tsay@chicagobooth.edu

In recent years, an extensive literature has focused on the L1 penalized least squares (Lasso) estimators of high dimensional linear regression when the number of covariates p is considerably larger than the sample size n. However, there is limited attention paid to the properties of the estimators when the errors or the covariates are serially dependent. In this study, we investigate the theoretical properties of the Lasso estimators for linear regression with random design under serially dependent and/or non-sub-Gaussian errors and covariates. In contrast to the traditional case where the errors are i.i.d and have finite exponential moments, we allow for p to be at most a power of n if the errors have only polynomial moments. In addition, the rate of convergence becomes slower due to the serial dependencies in errors and the covariates. We also consider sign consistency for model selection based on Lasso. Adopting the framework of functional dependence measure, we provide a detailed description on how the rates of convergence and the selection consistency of the estimators depend on the dependence measures and moment conditions of the errors and the covariates.

Applications to Nowcasting in which serially correlated errors and a large number of covariates are common

Model selection for high-dimensional sparse nonlinear models using Chebyshev greedy algorithms

Ching-Kang Ing¹, You-Lin Chen² and Wei Biao Wu²
¹ National Tsing Hua University
² University of Chicago
cking@stat.nthu.edu.tw

We are concerned with model selection problems in high-dimensional sparse nonlinear models. We first use the Chebyshev greedy algorithm (CGA) to perform variable screening and derive, under a fairly general sparsity condition, its rate of convergence in terms of the number of iterations and the approximation error. We then introduce a high-dimensional information criterion (HDIC) to determine the number of CGA iterations and show that CGA used in conjunction with HDIC achieves the optimal rate of convergence. Finally, the proposed method is applied to the analysis of high-dimensional logistic regression models.

Classical Backfitting for Smooth-backfitting Additive Models

Li-Shan Huang and Chung-Hsin Yu
National Tsing Hua University, TAIWAN
lihuang@stat.nthu.edu.tw

Smooth backfitting has been shown to have better theoretical properties than classical backfitting for fitting additive models based on local linear regression (Mammen et al., 1999). In this paper we show that the smooth backfitting procedure in the local linear case can be alternatively performed as a classical backfitting procedure with a different type of smoother matrices; the two criterion functions differ by some constant while the iterative equations are identical. The connections allow the smooth backfitting algorithm to be understood in a much simplified way, give new insights on the differences between the two approaches in the literature, and provide an extension to local polynomial regression. In the case of bivariate additive models, allowing for different orders of local polynomials and different bandwidths, asymptotic properties are investigated and cases of oracle efficiency are discussed. Simulations are conducted to demonstrate finite sample behaviors of the methodology.

Online Learning with Applications to Communication Network Traffic Management

Henry Horng-Shing Lu
National Chiao Tung University, Taiwan
hselu@stat.nctu.edu.tw

Supervised learning based on the methods of support vector machine (SVM) and the related techniques are very useful for the classification of complex data. However, the computation cost is very high when the training data set is massive. Online learning problems will need to handle the problems of memory limitation and computational complexity. In this study, the online learning meth-
ods by SVM and the related techniques for multi-class problems in massive data are developed. The empirical performance of these methods will be evaluated by real data in communication network traffic management.

**Session 148: Recent Developments in Theory and Application of Multiple Comparison Methods**

A Gatekeeping Test on a Primary and a Secondary Endpoint in a Group Sequential Design with Multiple

♦ Ajit Tamhane†, Jiangtao Gou‡, Christopher Jennison§, Cyrus Mehta¶ and Teresa Curto∥

1 Northwestern University
2 Hunter College
3 University of Bath
4 Cytel, Inc.

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 preassigned 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 O'Brien-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)).

**Multiplicity issues in clinical trials from a regulatory perspective**

Yongman Kim

FDA

During the last 10 years, there has been a big advancement in statistical approaches to handle multiplicity problems arising in clinical trials, mainly driven by academics and the pharmaceutical industry. Early this year, in order to fulfill a commitment to the FDA Amendment Act of 2007, FDA issued a draft guidance on multiplicity issues with the current regulatory thinking for public comments. The purpose of the guidance is to describe various strategies for grouping and ordering multiple endpoints and applying statistical methods for managing multiplicity within a study in order to control the chance of making erroneous conclusions about a drug’s effects. In this presentation, in line with the draft guidance, the main focus is on the multiplicity issues due to the multiple endpoints common in registration trials aiming for approval of new drugs and inclusion of clinically relevant claims in the product label. General principles and selected statistical methods that appear in the draft guidance are discussed. These include, for example, regulatory guidance for strong control of the type I error probability, understanding of alpha recycling and graphical approaches, and recommendations for gatekeeping strategies, etc. Regulatory issues related to labeling in association with multiple endpoints are also discussed.

**Lossless type-I error rate control for multiplicity and its application**

Muhammad Jalaluddin

AbbVie

Multiple statistical tests are common in clinical trials. There are widely used methods to control familywise type-I error rate. Many of these methods are based on the Bonferroni procedure. The Bonferroni procedure for multiplicity does not utilize the entire preassigned overall type-I error rate. Therefore, Bonferroni procedure based methods for multiplicity may also underutilize preassigned overall type-I error rate. A lossless procedure for multiplicity is introduced, and its application is demonstrated.

**Partition to Power Subgroup Inference in Personalized Medicine**

Szu-Yu Tang†, Yi Liu‡ and Jason Hsu§

1 Ventana Medical Systems, Inc.
2 Takeda Pharmaceuticals International Co.
3 The Ohio State University

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 which patients to target, inference on efficacy of the drug on the g+ patients, on the g- patients, and on the mixture g+, g- patients are needed. There are logical relationships among efficacy parameters in g+, g-, and g+, g-, which intersection hypotheses formed by Closed Testing would not recognize, but partitioning hypotheses formed by the Partition Principle would. This presentation first shows that Hazard Ratio and Odds Ratio are not Subgroup Mixable efficacy measures, but difference of means, ratio of response rates, and ratio of medians are. It then shows, for Subgroup Mixable efficacy measures, by recognizing logical relationships among the efficacy parameters, Partition Testing does not leave alpha on the table, and therefore empowers subgroup inference in personalized medicine.

**Session 149: Go/No-Go decision making in clinical trials**

**Enhancing the Probability of Success Framework for Go/No-Go Decision Making Using ROC Curves**

♦ Qi Tang†, Alan Hartford‡, Deli Wang§, Dey Jytirimoy¶, Walt Offen∥ and Frank Shen¶

1 Translational Informatics, Sanofi, Bridgewater, NJ
2 Statistical Innovation, AbbVie, Chicago, IL
3 Pipeline Statistics, AbbVie, Chicago, IL
4 Statistical Sciences, AbbVie, Chicago, IL
5 Data and Statistical Sciences, AbbVie, Chicago, IL

To make an informed Go/No Go decision for an investigational product, all available information should be effectively used to predict the results for ongoing and/or future clinical trials. This work is motivated by Broglio et al. (2014 Trials), which presented an interesting application of assessing probability of success in a business acquisition case study involving a key oncology compound in phase 3 development. We propose a similar framework for Go/No Go decision making as Broglio where probability of success is still assessed utilizing all relevant information in a Bayesian way and sensitivity of probability of success is still examined against different choices of priors. However, the methodology is enhanced by
including an evaluation of the risks associated with each decision made based using ROC curves. This enhancement of the methodology provides a quantitative approach to guide decision makers as to how high the probability of success should be to make a Go decision. A due diligence example is given to illustrate this enhanced framework.

**Interim Go/No-Go decision making in a phase 2 clinical trial of a neurodegenerative disorder**

*Weining Robieson, Greg Cicconetti and Deli Wang*

1AbbVie

weining.z.robieson@abbvie.com

Interim Go/No-Go decision making in a phase 2 clinical trial of a neurodegenerative disorder, Go/No-Go decisions at interim analyses (IAs) are a challenging part of the study design. It was not ideal to use only data at final visit for IAs due to the long treatment duration, which means few patients will have completed the trial when the enrollment ends. This special feature motivated consideration of earlier visits at IAs. We evaluated a mixed model repeated measure analysis at each IA using observed data and imputed data for future visits for ongoing subjects based on a linear disease progression assumption. The imputed Week 52 effect size was chosen as the metric for interim Go/No-Go criteria. Thresholds for imputed effect sizes were defined to satisfy pre-specified criteria for operating characteristics under different simulated longitudinal response scenarios. Go/No-Go decisions at IA will be based on whether the imputed effect size exceeds the chosen thresholds. We compared the performance of this linear imputation approach with no imputation under various IA timing considerations.

**Statistical Software for Decision Making in Clinical Development**

*Charles Liu*

Cytel

charles.liu@cytel.com

Lalonde et al. (2007) proposed a model-based approach to drug development, which emphasized quantitative criteria and trial performance metrics for Go/No-Go decisions. This talk will outline some of the major advantages of this approach, including the flexibility to use multiple decision criteria within a frequentist or Bayesian framework, and generalization to multiple interim looks. Issues of criteria-setting and optimization of metrics will be discussed with examples using statistical software.

**Session 150: Modern Advancements in Semi-/Non-parametric Statistics**

**Free-knot spline for Generalized Regression Models**

*Xinyi Li, Li Wang and Dan Nettleton*

1Iowa State University

li.xinyi@iastate.edu

2Iowa State University

xsong@uga.edu

The additive partially linear model (APLM) combines the flexibility of nonparametric regression with the parsimony of regression models, and can deal with high-dimensional data by avoiding “curse of dimensionality”. A natural question raised in practice is the choice of structure of the nonparametric part, that is, whether the quantitative covariates enter into model in linear or nonparametric form. We present a comprehensive framework for simultaneous sparse model identification and learning for APLMs. For this purpose, we propose an efficient two-stage procedure for ultra-high-dimensional APLMs where both the linear and nonparametric components are possibly larger than the sample size. In the first stage, we decompose the nonparametric functions into linear and truncated nonlinear functions. The truncated nonlinear functions are approximated by constant spline bases, and a triple penalization procedure is proposed to select nonzero components using adaptive group LASSO. In the second stage, we refit data with selected covariates, and apply local linear smoothing to obtain the asymptotic normality for the estimators. The procedure is shown to be consistent for model structure identification. It can identify zero, linear, and nonlinear components correctly and efficiently. Inference can be made on both linear coefficients and nonparametric functions. We conduct simulation studies to evaluate the performance of the method, and apply the proposed method to a dataset on the Shoot Apical Meristem (SAM) of maize genotypes for illustration.

**Partially time-varying coefficient proportional hazards models with error-prone covariates**

*Xiao Song and Li Wang*

1University of Georgia

2Iowa State University

Due to cost and time considerations, interest has focused on identifying surrogate markers that could be substituted for the clinical endpoint, time to an event of interest, in evaluation of treatment efficacy. Joint models are often used to assess the effect of surrogate markers and treatment. Motivated by recent works studying the AIDS Clinical Trial Group (ACTG) 175 data, we propose a partially time-varying coefficient proportional hazards model for modeling the relationship between the hazard of failure and time-dependent and time-independent covariates. The time-varying coefficients are approximated by polynomial splines, and the corrected score and conditional score approaches are adopted to estimate the regression coefficients. The proposed estimators are consistent, and the asymptotic normality is established for the constant coefficients, which enables us to construct confidence intervals and permits joint inference. The finite-sample performance of the proposed method is assessed by Monte Carlo simulation studies. The proposed model is applied to ACTG 175 data to assess the temporal dynamics of the effect of treatment and CD4 count on time to AIDS or death.
Session 151: Modern Statistical methods for Educational Testing

Bayesian estimation of the DINA Q matrix

Yinghan Chen and Steve Culpepper
University of Illinois at Urbana-Champaign

Cognitive diagnosis models are partially ordered latent class models and are used to classify students into skill mastery profiles. The deterministic inputs, noisy “and” gate model (DINA) is a popular psychometric model for cognitive diagnosis. Application of the DINA model requires content knowledge of a Q matrix, which maps the attributes or skills needed to master a collection of items. Misprediction of Q has been shown to yield biased diagnostic classifications. We propose a Bayesian framework for estimating the DINA Q matrix. The developed algorithm builds upon prior research (Chen, Liu, Xu, & Ying, 2015) and ensures the estimated Q matrix is identified. Monte Carlo evidence is presented to support the accuracy of parameter recovery. The developed methodology is applied to Tatsuoka’s fraction-subtraction dataset.

Latent Class Models for Learning

Shiyu Wang, Steve Culpepper, Yinghan Chen and Jeff Douglas

Latent class models for learning are presented in the context of a spatial reasoning example. A conjunctive cognitive diagnosis model is combined with a transition model for learning that utilizes time-dependent covariates as well as a latent learning model parameter. A variety of other learning models are discussed and the notion of item-specific learning parameters is presented. The use of response times as indicators of both learning and fluency is considered as a future development.

A Fused Latent and Graphical Model

Jingchen Liu
Columbia University

One of the main tasks of statistical models is to characterize the dependence structures of multi-dimensional distributions. Latent variable models, especially when the dimension gets small, are useful for modeling such dependence structures. A popular approach is to model the dependence of the latent variable(s) on the observed data. In this paper, a novel fused latent and graphical model is proposed. The model is based on a functional factor model and the sparse precision matrix is used to characterize the dependence structure. The model is shown to be useful for the analysis of high-dimensional data.

Practical Considerations for Educational Testing

Betsy McCoach
University of Connecticut

Recent technological and statistical advances allow for greater flexibility and sophistication within the test development process. Increasingly, test users expect more nuanced and thorough information and interpretation. To what degree can modern psychometrics aid in solving issues related to student learning and educational accountability? This talk will discuss the practical issues that test developers and test users should consider in this era where increasingly technical tests are met with rising skepticism from educators and the general public.

Session 152: Variable Selection and Dimension Reduction for High-Dimensional Models

Linear and Nonlinear Sufficient Dimension Reduction for Functional Data

Bing Li and Jun Song

When you sign your name on a touch-pad device, your handwriting is recorded as a parameterized curve along with the time. GPS sensors can record a person’s movement in three-dimensional space as a function of time. These types of data can be considered as a realization of a random multivariate function. They are increasingly prevalent in modern applications and their analysis intrinsically requires probability model over functional space of infinite dimensions. In this talk, we will discuss sufficient dimension reduction (SDR) for functional data in which both predictor and response are multivariate functional data. First, we introduce nested reproducing kernel Hilbert space (nested RKHS) which provides a general mechanism for nonlinear functional data analysis. Two layers of function spaces are constructed in a nested fashion so that the first space represents the observed functional data, and the second space characterizes nonlinearity of the random functions. As well as nonlinear functional data analysis, the nested RKHS enables us to develop linear SDR for multivariate functional data without using a slice of the response. Then we will introduce a method of nonlinear SDR for functional data, the detailed procedure for estimation, and its consistency and convergence rate. The applications to speech recognition and handwriting symbol association show that the reduced predictors can be used for classification with great effectiveness.

Joint mean and covariance estimation based on model selection with unreplicated matrix-variate data

Michael Hornstein, Roger Fan, Kerby Shedden and Shuheng Zhou
University of Michigan

In the setting of matrix-variate data, dependencies exist among both observations and variables, resulting in the challenging problem of analyzing unreplicated high-dimensional data with unknown mean and dependence structures. We present a method for jointly estimating the mean and covariance, using penalized inverse covariance estimation and generalized least squares. The method uses a preliminary model selection step to remove the mean structure before estimating the covariance. Our approach allows more efficient estimation of the mean structure by accounting for the dependence structure. Furthermore, inferences about the mean parameters become correctly calibrated. We present rates of convergence for estimating the mean parameters and covariance matrices, and we illustrate the performance of the procedure using simulation results and an analysis of genomic data.

Generalized Principal Component Analysis

Yoonkyung Lee and Andrew Landgraf

Principal component analysis (PCA) is useful for a wide range of data analysis tasks. However, its implicit link to the Gaussian distribution can be undesirable for discrete data such as binary and multiclass responses or counts. We generalize PCA to handle various types of data using the generalized linear model framework. In contrast to the existing approach of matrix factorizations for exponen-
Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications
Xiaohui Chen
University of Illinois at Urbana-Champaign

This paper studies the Gaussian and bootstrap approximations for the probabilities of a non-degenerate U-statistic belonging to the hyperrectangles in $\mathbb{R}^d$ when the dimension $d$ is large. A two-step Gaussian approximation procedure that does not impose structural assumptions on the data distribution is proposed. Subject to mild moment conditions on the kernel, we establish the explicit rate of convergence uniformly in the class of all hyperrectangles in $\mathbb{R}^d$ that decays polynomially in sample size for a high-dimensional scaling limit, where the dimension can be much larger than the sample size. We also provide computable approximation methods for the quantiles of the maxima of centered U-statistics. Specifically, we provide a unified perspective for the empirical bootstrap, the randomly reweighted bootstrap, and the Gaussian multiplier bootstrap with the jackknife estimator of covariance matrix as randomly reweighted quadratic forms and we establish their validity. We show that all three methods are inferentially first-order equivalent for high-dimensional U-statistics in the sense that they achieve the same uniform rate of convergence over all $d$-dimensional hyperrectangles. In particular, they are asymptotically valid when the dimension $d$ can be as large as $O(e^{c_n})$ for some constant $c \in (0, 1/7)$.

The bootstrap methods are applied to statistical applications for high-dimensional non-Gaussian data including: (i) principled and data-dependent tuning parameter selection for regularized estimation of the covariance matrix and its related functionals; (ii) simultaneous inference for the covariance and rank correlation matrices. In particular, for the thresholded covariance matrix estimator with the bootstrap selected tuning parameter, we show that for a class of subgaussian data, error bounds of the bootstrapped thresholded covariance matrix estimator can be much tighter than those of the minimax estimator with a universal threshold. In addition, we also show that the Gaussian-like convergence rates can be achieved for heavy-tailed data, which are less conservative than those obtained by the Bonferroni technique that ignores the dependency in the underlying data distribution.

Session 153: Advances in joint models for longitudinal and time-to-event data with biomedical applications

Informative Cluster Size and Observation Time in Longitudinal Transition Models
Joe Bible, Danping Liu and Paul Albert
National Institutes of Health
danping.liu@nih.gov

Transition models are useful in estimating the probability of recurrent events in longitudinal studies. However, direct application of a transition model may suffer from two complications, informative cluster size and informative gap time between observations. For example, Consecutive Pregnancy Study (CPS) is a retrospective cohort study aiming at understanding the recurrence patterns and predictors of adverse pregnancy outcomes, such as preterm birth. The number of pregnancies observed and the gap time may be both indicative of a women’s underlying fertility, and hence correlated with the pregnancy outcomes. We propose a shared random effect structure for jointly modeling the transition model with the informative observation process. The gap time is modeled by a parametric distribution with right censoring; the cluster size is characterized by a continuation ratio model. We also investigated the estimation and interpretation of two transition probabilities: one adjusted for gap time and the other marginalized over gap time. Through extensive simulation studies and analyses of the CPS data, we showed that naive approaches ignoring the informative observation process could lead to seriously biased inference.

Statistical Methods for Multivariate Failure Time Data Analysis
Shanshan Zhao\textsuperscript{1} and Ross Prentice\textsuperscript{2}
\textsuperscript{1}NIEHS/NIH
\textsuperscript{2}Fred Hutchinson Cancer Research Center
shanshan.zhao@nih.gov

In many epidemiology studies, participants are followed for more than one health outcome. For example, women can develop related diseases (e.g., breast cancer and ovarian cancer), competing diseases (e.g., breast cancer and death), and recurrent diseases (e.g., multiple breast cancer onsets). Although univariate survival data analysis has been well developed, methods for multivariate survival data analysis are limited. Most studies analyzed these outcomes separately, or adjusted status of other diseases as covariates in the risk model of the main disease of interest. We propose to model the joint risks of multiple diseases through marginal Cox proportional hazards models and a Cox-type model for the cross-ratio. Regression parameters can be estimated by maximizing the profile likelihood. The proposed method is applied to the Women’s Health Initiative’s hormone therapy trial data, and it revealed a positive dependency between coronary heart disease and stroke, and also the dependency is stronger for women under hormone therapy.

A Frailty-Copula Approach for Modeling a Terminal Event and Recurrent Events with Biomarker Data
Zheng Li\textsuperscript{1} and Ming Wang\textsuperscript{2}
\textsuperscript{1}Vernon Chinchilli
\textsuperscript{2}Pennsylvania State University
zxil4@psu.edu

Recurrent events sometimes will be censored by a terminal event. The non-informative censoring assumption is violated under this situation, and as a result, we cannot model the recurrent event process alone. The joint frailty model is widely used to jointly model these two processes. However, under this model, we assume the terminal event process and the recurrent event process are conditionally independent given a subject-level frailty. Also, we cannot directly estimate the association between two time-to-event processes. In order to relax the conditional independence assumption and estimate the association directly, we propose a joint frailty-copula approach under a Bayesian framework to jointly model the terminal time-to-event process and the recurrent time-to-event process. We further incorporate a longitudinal biomarker process into the bayesian joint frailty-copula model. We show that the joint frailty-copula model is a more generalized model, extended from a nested frailty model. Metropolis-Hastings within Gibss Sampler algorithm is used for estimating parameters. Extensive simulation studies are performed to evaluate the performance of our method in terms of bias, mean...
Joint models of longitudinal outcome and interval-censored events with time varying covariate effect

Bin Zhang and Yue Zhang
Cincinnati Children’s Hospital Medical Center and Shanghai Jiaotong University

Joint models of survival and longitudinal data have gained more and more attention in the past several decades. However, only limited methods are developed for interval-censored time to event data. In this paper, we develop a Bayesian approach for joint models of longitudinal and interval censored survival data. Time varying coefficients are considered when temporal effect exists. The correlation between longitudinal markers and the time to events is assessed by latent variables. The properties of the proposed model are illustrated by simulation studies. An application to a national stroke study is conducted to demonstrate the performance of the proposed method.

Session 154: Recent Advances on High-Dimensional Statistics

Simultaneous Multistage Adaptive Ranking and Thresholding for Sparse Signal Recovery
Wenguang Sun and Weinan Wang
University of Southern California

Multistage design provides a cost-effective way to glean significance from data by adaptively reducing a large set of variables to a much smaller set in a sequential manner. By allocating sensing resources adaptively according to the estimated superset of the signal support, one can effectively eliminate null locations and localize signals with a much smaller study budget. We formulate a compound decision theoretical framework for simultaneous multistage adaptive testing and study how to minimize the total number of measurements while meeting pre-specified constraints on both the false discovery rate and missed discovery rate. The new procedure, which effectively pools information from all stage-wise tests using a simultaneous multistage adaptive ranking and thresholding (SMART) scheme, and can greatly reduce the approximation errors in conventional individual sequential probability ratio tests and hence leads to great savings in total sensing efforts. It is shown that SMART controls the FDR and MDR at the nominal levels and is optimal in the sense that it achieves the information theoretic lower bounds. Numerical studies confirm the validity of SMART for FDR and MDR control and show it achieves substantial power gain over other existing methods. The SMART procedure is demonstrated through the analysis of spatial imaging data.

Community Detection in Sparse Networks Using the Symmetrized Laplacian Inverse Matrix (SLIM)
Bingyi Jing, Ting Li, Ningchen Ying and Xianshi Yu

Previous work on network community detection focused mainly on dense scenarios, where the average degree of nodes $E(\text{degree}) = \Omega(\log n)$. Recently, sparse networks have received increasing attention amongst researchers. In this paper, we propose a new method for detection communities in sparse networks by using the symmetrized Laplacian inverse matrix (SLIM) to measure the closeness between nodes. The idea comes from the first hitting time in random walk, and it also has a nice interpretation in diffusion maps. Community membership is acquired by applying the spectral method to the SLIM. We show that the SLIM is asymptotically consistent for the stochastic block model (SBM) as sparse as $E(\text{degree}) \rightarrow \infty$. The SLIM outperforms other methods in many simulations and real datasets. It performs very well for the degree-corrected SBM. The computational complexity of the SLIM is $O(n^3)$, which can be further reduced to $O(n^2)$ by a finite power matrix series.

Spatial Adaptation in Trend Filtering
Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee and Bodhisattva Sen
University of California, Berkeley and University of Chicago

We study trend filtering, a relatively recent method for univariate nonparametric regression. For a given integer $r \geq 1$, the trend filtering estimator of order $r$ is defined as the minimizer of the sum of squared errors when we constrain (or penalize) the sum of the absolute discrete derivatives of order $r$ over the input points. For $r = 1$, the estimator reduces to total variation regularization. For a given integer $r \geq 1$, the trend filtering estimator of order $r$ is defined as the minimizer of the sum of squared errors when we constrain (or penalize) the sum of the absolute discrete derivatives of order $r$ over the input points. For $r = 1$, the estimator reduces to total variation regularization which has received much attention in the statistics and image processing literature. In this paper, we study the performance of the trend filtering estimator for every $r > 1$, both in the constrained and penalized forms. Our main results show that the estimator is optimally spatially adaptive. Specifically, we prove that, in spite of being a nonparametric estimator, when the underlying function is a (discrete) spline with few "knots", the risk (under the global squared error loss) of the trend filtering estimator (with an appropriate choice of the tuning parameter) achieves the parametric $n^{-1}$-rate, up to a logarithmic (multiplicative) factor. Our results also include high probability statements on the loss and are stated in sharp oracle form, i.e., our results apply to misspecification and have leading constant one for the misspecified term. Moreover, some of the metric entropy results used in this paper, derived using connections to fat shattering, are of independent interest.

Session 155: Utilizing External Data in a Medical Product Evaluation Study: Statistical Design and Analysis Consid...
impact of violation of the constancy assumption on the validity of a non-inferiority trial: prevention through good trial conduct and adjustment through analysis. In this presentation, we will propose a new approach through including a companion constancy test to ensure the main non-inferiority test so consequently to protect the validity of a non-inferiority trial. We will use an example to illustrate our approach.

On Weighted Performance Goals in Medical Device Single-Arm Clinical Studies

Nelson Lu, Wei-Chen Chen, Yunling Xu and Ram Tiwari
FDA/CDRH
nelson.lu@fda.hhs.gov

Under some circumstances, a pre-market application of medical devices may be supported by a single arm study where clinical endpoint is compared with a performance goal (PG). Occasionally, performance is expected to be significantly different between two clinically defined subgroups of patients. For some products, it has been accepted to compare the combined results with a weighted PG. Some methods under the frequentist framework have been developed. In this presentation, we propose to analyze such data under the Bayesian framework.

A Study Design for Utilizing External Data to Augment the Control in a Randomized Controlled Trial

Yunling Xu, Nelson Lu, Lilly Yue and Ram Tiwari
FDA/CDRH
yun-ling.xu@fda.hhs.gov

With the increasing demand for utilization of historical study data and registry data to reduce the time and cost of bringing a medical product to the market, study validity and integrity is among the statistical challenges that need to be carefully addressed. We discuss a study design process to enhance the study validity and integrity for augmenting the control group in a randomized controlled trial utilizing external data. The discussed design process is tailored to confirmatory purpose emphasizing sequential access of data by purpose, and it covers selection of patients from historical studies/registries and assessment of similarity in patient characteristics between the current study and the selected control group; and assessment of similarity in patient clinical outcome between the current study control group and the selected control group is also discussed.

Session 156: Integrative methods for challenging genomics data

Clustering with Hidden Markov Model on Variable Blocks

Lynn Lin
Pennsylvania State University
llin@psu.edu

Advances in single-cell technologies have enabled high-dimensional measurement of individual cells in a high-throughput manner. A key first step to analyze this wealth of data is to identify distinct cell subsets from a mixed-population sample such as blood or tissue. In many clinical applications, cell subsets of interest are often found in very low frequencies. This poses challenges for existing clustering methods. To address this issue, we propose a new mixture model, namely the Hidden Markov Model on Variable Blocks (HMM-VB). HMM-VB incorporates prior knowledge on the chain-like dependence structure among groups of variables, achieving the effect of dimension reduction as well as incisive modeling of the rare clusters. In a series of experiments on simulated and real data, HMM-VB outperforms other widely used methods.

A hierarchical hidden Markov random field model for peak calling across multiple Hi-C datasets

Joshua Martin1, Ming Hu2 and Yun Li3
1 University of North Carolina
2 Cleveland Clinic Foundation
jsmartin@email.unc.edu

The constantly accumulating Hi-C data provide rich information for calling peaks across multiple tissue/cell types, experimental conditions, and/or cell differentiation stages. However, statistical models and computational tools are still in their infancy. Multiple factors, including sequencing depth and heterogeneity across Hi-C experiments, pose great challenges for the development of proper and efficient methods. We propose a peak caller based on a hierarchical hidden Markov random field model we call MuNIn (Multi-tissue Nucleome Investigator) to detect long range chromatin interactions from multiple Hi-C datasets. In addition to model the spatial dependency in the local neighborhood (Xu et al., Bioinformatics, 2015 & 2016), MuNIn is able to model dependency across multiple Hi-C datasets, leading to further improved statistical power. We conducted comprehensive simulation studies, and showed that MuNIn outperforms competing methods that ignore the dependency structure and call peaks separately in each individual Hi-C dataset. Next, we analyzed a real Hi-C dataset on human H1 embryonic stem cells and four H1 derived cells (Xie et al., Cell, 2013), and found that the cell-type-specific peaks identified by MuNIn show higher overlap with cell-type-specific epigenetic features and cell-type-specific gene expression, compared to those identified by competing methods. MuNIn has the potential to unveil the structural basis of cell-type-specific transcription mechanism.

Bayesian Approaches to Integrative Genomic Analysis: Data Integration and Scalable Computing

Xiaoquan Wen
University of Michigan
xwen@umich.edu

With the increasing availability of functional genomic data, incorporating genomic annotations into genetic association analysis has become an increasing trend in the field of genetics and genomics. However, the existing methods often lack rigor and/or computational efficiency and consequently do not maximize the utility of genomic annotations. In this talk, we discuss a rigorous Bayesian inference procedure to perform integrative analysis of genomic data for both traditional genome-wide association analysis (GWAS) and emerging molecular QTL mapping studies. In particular, we show that Bayesian strategies present some unique advantages in naturally integrating multiple sources of data and framing the underlying scientific questions into model comparison/selection problems. Taking advantages of the context-specific domain knowledge, we develop a set of computational methods to efficiently make inference. We demonstrate our approaches by analyzing a multi-tissue expression quantitative trait loci (eQTL) data sets from the NIH GTEx project and GWAS data from Global lipids consortium.

Assessing reproducibility of Hi-C chromatin loops using irreproducible discovery rate regression

Feipeng Zhang, Tao Yang and Qunhua Li
Penn State University
qunhua.li@psu.edu

High-throughput chromosome conformation capture (Hi-C) has become the state-of-the-art technology for studying genome-wide chromatin structures. Despite rapid technological advance, Hi-C data still is highly noisy. Assessing reproducibility of Hi-C data across replicates is an important way to monitor data quality.
Though some statistical methods have been developed for assessing reproducibility of findings from high-throughput experiments represented in the form of ranked lists, these methods are not entirely suitable for Hi-C data, as Hi-C data has a strong spatial structure that is not taken account of in the existing methods.

In this work, we develop the irreproducible discovery rate regression, a method for incorporating covariate information into the assessment of reproducibility. By modeling the spatial structure through covariates, this method effectively takes account of the spatial effect in Hi-C data and improve the identification of reproducible signals. In fact, this method is generic and is applicable to many other settings. We illustrate our method using both simulations and real data analyses.

Session 157: Functional Data Analysis in Biomedical Research

Dynamic Modeling of Longitudinal Snippets

Hans-Georg Mueller and Matthew Dawson

University of California, Davis

hgmuller@ucdavis.edu

Longitudinal data are often plagued with sparsity of time points where measurements are available. The functional data analysis perspective provides an effective and flexible approach to address this problem for the commonly studied case where measurements are sparse but their times are randomly distributed over an interval. Here we consider a different scenario of data snippets. These are very short stretches of longitudinal measurements, which arise in “accelerated” longitudinal studies. For each subject the stretch of available data is much shorter than the time frame of interest. An added challenge is introduced if a meaningful time proxy such as time since disease onset is not available. We approach this problem through conditional quantile trajectories for monotonic processes that arise as solutions of a dynamic system and discuss consistent estimates and an application to shrinking brain volumes in Alzheimer’s patients.

Optimal Penalized Function-on-Function Regression under a Reproducing Kernel Hilbert Space Framework

Xiaoxiao Sun1, Pang Du2, Xiao Wang3 and Ping Ma1

1University of Georgia
2Virginia Tech
3Purdue University

xiaosun@uga.edu

Many studies collect data with response and predictor variables both being functions of some covariate. Their common goal is to understand the relationship between these functional variables. Motivated by two real-life examples, we propose a new function-on-function regression model that can be used to analyze such kind of functional data. Our estimator of the 2D coefficient function is the optimizer of a form of penalized least squares where the penalty enforces a certain level of smoothness on the estimator. Our first result is the Representer Theorem which states that the exact optimizer of the penalized least squares actually resides in a data-adaptive finite dimensional subspace although the optimization problem is defined on a function space of infinite dimensions. This theorem then allows us an easy incorporation of the Gaussian quadrature into the optimization of the penalized least squares, which can be carried out through standard numerical procedures. We also show that our estimator achieves the minimax convergence rate in mean prediction under the framework of function-on-function regression.

Registration for Binary Functional Data

Julia Wrobel and Jeff Goldsmith

Columbia University

jw3134@columbia.edu

We consider the problem of aligning curves from exponential family distributions. The approach is based on the combination of alignment and functional principal components analysis, and is facilitated by recent extensions of FPCA to non-Gaussian settings. Our work is motivated by the study of physical activity using accelerometers, wearable devices that provide around-the-clock monitoring of activity and produce non-Gaussian measurements. We apply the proposed methods to activity counts to a binary “active” vs “inactive” indicator using a binomial distribution. After alignment, the trajectories show clear peaks of activity in the morning and afternoon with a dip in the middle of the day.

Session 158: Urging a paradigm change: New developments on statistical inferences

Approximate CD Computing: An effective likelihood-free method with statistical guarantees

Suzanne Thornton and Min-ge Xie

Rutgers, The Stat University of New Jersey

suzanne.thornton@rutgers.edu

Approximate Bayesian computing (ABC) is a likelihood-free method that has grown increasingly popular since early applications in population genetics. However, the theoretical justification for inference based on this method has yet to be fully developed especially pertaining to the use of non-sufficient summary statistics. We introduce a more general computational technique, approximate confidence distribution computing (ACC) to overcome a few issues associated with the ABC method, for instance, the lack of theory supporting the use of non-sufficient summary statistics, the lack of guardian for the selection of prior, and the long computing time. Specifically, we establish frequentist coverage properties for the outcome of the ACC method by using the theory of confidence distributions, and thus inference based on ACC is justified, even if reliant upon a non-sufficient summary statistic. Furthermore, the ACC method is very broadly applicable; in fact, the ABC algorithm can be viewed as a special case of an ACC method without damaging the integrity of ACC based inference. We supplement the theory with simulation studies and an epidemiological application to illustrate the benefits of the ACC method. It is demonstrated that a well-tended ACC algorithm can greatly increase its computing efficiency over a typical ABC algorithm.

Partial Bayes: Exact Inference with Partially Specified Bayesian Models

Yixuan Qiu, Lingsong Zhang and Chuanhai Liu

Purdue University

yixuanq@gmail.com

Bayesian methods are useful for statistical inference. However, real-world problems can be challenging when using Bayesian methods when the data analyst has only limited prior knowledge. In this talk we consider a class of partially specified Bayesian models that include those that were previously handled by the Empirical Bayes method. Taking the recently proposed Inferential Model approach, we develop methods for both exact and efficient solutions for a class of commonly used hierarchical models. We also discuss an important case where prior distributions can only be available for a lower-dimensional function of the parameters. In addition to the theoretical investigation, numerical results and real applications are used.
to demonstrate the superior performance of the proposed method, called the Partial Bayes method.

**Personalism and Dempster-Shafer Analysis for the 21st Century**

Paul Edlefsen¹, Ruobin Gong² and Arthur P. Dempster²

¹Fred Hutchinson Cancer Research Center
²Harvard University
pedelefs@fredhutch.org

Personalism, like its more familiar cousins objectivism and subjectivism, is a perspective on the role of the statistician (or more generally, the scientist: “you”) in the conduct of science. The differences among these perspectives usually impact little on the daily conduct of statisticians, to our great relief. We are already sufficiently aware of our expected role on a paper or grant to ensure that the science is reproducible. However we are also not unaware of the several difficult issues that we face in statistical science including our contribution to the too-often disappointing performance of phase 3 trials, which results in a high cost for the biomedical enterprise and possibly in lost opportunity for quality human living. There are a host of thorny issues we must face, including how to maximize information yield from research investments while accounting for multiple testing and post-hoc inference. Furthermore, “p ≫ n” problems (high dimensional covariates with low numbers of observations) arise increasingly often in clinical trials. Can “Fisher’s Greatest Blunder” (which is how Brad Efron described Fisher’s “Fiducial” methodology) possibly offer an insight into a solution to all of these problems? I will begin by introducing the personalist perspective on the role that “you” play in the scientific process and I will briefly describe Dempster-Shafer (DS) methodology for statistical inference and prediction. I will argue that DS potentially offers a path toward a truly cohesive statistical inference framework for the 21st century.

**Generalized Fiducial Inference for Nonparametric Function Estimation**

Randy Lai¹ and Raymond Wong²

¹University of Maine
²Iowa State University
chushing.lai@maine.edu

Generalized fiducial inference (GFI) is a relatively new approach for conducting statistical inference and performing uncertainty quantification, although its original form was introduced by Fisher back in the 1930s. In this project, we describe an alternative technique to investigate nonparametric function estimation problems in the GFI framework via reproducing kernel Hilbert space (RKHS).

**Session 159: Statistical Modeling of High Dimensional Data with Applications to Genomics**

**Modeling parent-of-origin effect in eQTL mapping using RNA-seq data**

Feifei Xiao¹, Shirong Deng², Guoshuai Cai³ and Christopher I. Amor³

¹University of South Carolina
²Augusta University
³Dartmouth College
xiao@mailbox.sc.edu

Genomic imprinting is an important epigenetic phenomenon where the expression of certain genes depends on their parent-of-origin. Many imprinting genes are known to play important roles in human complex diseases such as diabetes, breast cancer and obesity. In recent years, array based eQTL studies have identified many regulatory variants that show associations with gene expression level. Nowadays, the rapid arising next-generation-sequencing is often done for eQTL mapping. We believe that parent-of-origin effect can contribute to regulating gene expression along with the overall effect from the gene. However, multicollinearity occurs naturally when we are modeling multiple genetic components, such as additive, dominance and imprinting effects. Moreover, it has been repeatedly shown that RNA-seq data are over-dispersed which brings challenge to the modeling of the gene expression profiling. To address these issues, we introduced a novel method to test the main allelic effects along with the imprinting effect in detecting eQTL. We utilized an orthogonalization procedure, which allowed for efficient imprinting effect detection whereas maintained the power to detect the main allelic effect from eQTLs. We conducted extensive simulations to demonstrate the statistical behavior of the proposed method.

**A Deterministic Global Optimization Method for Variational Inference**

Hachem Saddiki¹, Andrew Trapp² and Patrick Flaherty¹

¹University of Massachusetts, Amherst
²Worcester Polytechnic Institute
fialherty@math.umass.edu

Variational inference methods for latent variable statistical models have gained popularity because they are relatively fast, can handle large data sets, and have deterministic convergence guarantees. However, in practice it is unclear whether the fixed point identified by the variational inference algorithm is a local or a global optimum. Here, we propose a method for constructing iterative optimization algorithms for variational inference problems that are guaranteed to converge to the ε-global variational lower bound on the log-likelihood. We derive inference algorithms for two variational approximations to a standard Bayesian Gaussian mixture model (BGMM). We present a minimal data set for empirically testing convergence and show that a variational inference algorithm frequently converges to a local optimum while our algorithm always converges to the globally optimal variational lower bound. We characterize the loss incurred by choosing a non-optimal variational approximation distribution suggesting that selection of the approximating variational distribution deserves as much attention as the selection of the original statistical model for a given data set. Finally, we describe applications of this strategy to genomic testing problems.

**Statistical framework for 3D chromatin structure modeling from Hi-C data**

Zhengqing Ouyang

The Jackson Laboratory for Genomic Medicine
zhengqing.ouyang@jax.org

Recent chromosome conformation capture technologies (such as Hi-C) are revolutionizing biomedical research by allowing three-dimensional (3D) characterization of the genome. While massive, complex data sets have been generated by 3D genomics technologies, few statistical approaches exist for effectively modeling the higher-level chromatin structure of the genome. Here, we introduce a model-based approach for reconstructing 3D chromatin structure from Hi-C data. Through simulation and real data applications, we demonstrate accurate and robust reconstruction of 3D chromatin structure at high resolution and genome-wide level.

**Pathway Lasso: Estimate and Select Causal Mediation Pathways with High Dimensional Mediators**

Yi Zhao and Xi Luo

Brown University
In many scientific studies, it becomes increasingly important to delineate the causal pathways through a large number of mediators, such as genetic and brain mediators. Structural equation modeling (SEM) is a popular technique to estimate the pathway effects, commonly expressed as products of coefficients. However, it becomes unstable to fit such models with high dimensional mediators, especially for a general setting where all the mediators are causally dependent but the exact causal relationships between them are unknown. This paper proposes a sparse mediation model using a regularized SEM approach, where sparsity here means that a small number of mediators have nonzero mediation effects between a treatment and an outcome. To address the model selection challenge, we innovate by introducing a new penalty called Pathway Lasso. This penalty function is a convex relaxation of the non-convex product function, and it enables a computationally tractable optimization criterion to estimate and select many pathway effects simultaneously. We develop a fast ADMM-type algorithm to compute the model parameters, and we show that the iterative updates can be expressed in closed form. Theoretical analysis demonstrates that our method is able to consistently estimate the mediation pathway effects. The proposed approach yields higher pathway selection accuracy and lower estimation bias than the competing methods in the simulation studies.
Index of Authors

Abecasis, G, 62, 148
Afendras, G, 47, 94
Agniel, D, 43, 80
Ahn, C, 65, 161
Ahn, M, 46, 91
Akakpo, R, 68, 174
Albert, P, 71, 181
Alekseyenko, A, 56, 125
Altzerinakou, M, 66, 164
Amos, CL, 72, 185
An, L, 49, 55, 101, 125
Arthur, V, 44, 83
Asakura, K, 45, 86
Aseltine, R, 50, 107
Atem, P, 48, 97
Aue, A, 58, 136
Baccarelli, A, 65, 158
Bacher, R, 55, 123
Bai, S, 47, 57, 63, 76, 130
Bai, X, 55, 123
Baladandayuthapani, V, 56, 127
Ballard, S, 42, 76
Ballman, K, 57, 131
Bally, M, 69, 174
Bandeen-Roche, K, 49, 104
Bar, H, 61, 145
Baro, E, 49, 104
Barron, M, 60, 140
Basu, C, 50, 105
Batzoglou, S, 55, 124
Begg, C, 56, 164
Begg, C, 57, 129
Belloc, PC, 66, 165
Bemis, K, 68, 172
Bergeman, C, 58, 133
Berger, J, 54, 121
Berry, D, 62, 147
Betensky, R, 48, 97
Betz, J, 48, 62, 98, 147
Bhadravati, A, 62, 150
Bi, X, 67, 168
Bible, J, 44, 181
Bien, J, 53, 116
Blaser, M, 49, 101
Bogdan, M, 42, 78
Bojinov, I, 47, 97
Booth, J, 61, 145
Bose, K, 42, 70
Bradie, J, 43, 80
Braun, T, 43, 82
Brazauskas, R, 63, 152
Bretz, F, 60, 63, 142, 154
Brynjarsdottir, J, 49, 102
Bunn, V, 46, 93
Burdick, R, 59, 139
Buyse, M, 57, 57, 130, 130
Caffo, B, 41, 75
Cai, G, 72, 185
Cai, J, 53, 68, 119, 171
Cai, L, 69, 176
Cai, X, 43, 53, 67, 80, 113
Cai, TT, 43, 83
Cair, X, 65, 66, 151, 163
Campbell, Gil, 65, 159
Canadas, E, 72, 78
Cao, C, 46, 93
Cao, H, 41, 73
Caporaso, N, 67, 167
Carlin, B, 50, 105
Carmichael, O, 51, 109
Carter, K, 55, 125
Carter, R, 53, 115
Carvalho, C, 45, 88
Caserta, M, 58, 136
Casoli, CD, 58, 132
Castruccio, S, 41, 74
Catenacci, D, 56, 93
Chai, H, 55, 125
Chai, X, 57, 128
Chai, Z, 45, 89
Chakraborty, R, 42, 78
Chambers, G, 65, 164
Chand, HP, 62, 150
Chang, J, 67, 166
Chang, KG, 64, 119
Chang, W, 63, 154
Chang, C, 46, 89
Chang, H, 51, 56, 58, 109, 125, 135
Chang, J, 43, 81
Chang, M, 66, 152
Chang, S, 59, 139
Chang, V, 66, 172
Chang, W, 41, 73
Chang, X, 52, 115
Chang, Y, 57, 129
Chattaraj, N, 59, 138
Chattaraj, S, 71, 182
Chaudhuri, S, 41, 75
Checkley, W, 58, 136
Chen, A, 58, 99
Chen, C, 51, 66, 67, 73, 127
Chen, D, 52, 149
Chen, G, 56, 125
Chen, H, 64, 66, 155, 163
Chen, HY, 52, 114
Chen, J, 51, 66, 57, 61, 61, 73, 90, 128, 145
Chen, K, 50, 51, 107, 110
Chen, L, 41, 55, 59, 73, 123
Chen, M, 42, 53, 61, 77
Chen, R, 59, 60, 104
Chen, S, 41, 47, 49, 61, 75
Cheung, YK, 42, 45, 46, 79, 139
Cheung, K, 53, 97
Cheung, K, 59, 103, 103
Cheung, M, 69, 87, 100, 108
Cheung, YK, 42, 45, 46, 79, 139
Cheng, A, 59, 139
Cheng, C, 63, 153
Cheng, G, 48, 60, 64, 98
Cheng, J, 54, 122
Chenouri, S, 53, 117
Cheung, K, 43, 25, 82, 89
Cheung, YK, 42, 45, 46, 79
Chi, E, 67, 168
Chi, Y, 44, 85
Chiang, A, 60, 142
Chiang, A, 57, 129
Choiou, SH, 58, 119
Chiu, C, 43, 81
Chiu, C, 45, 89
Cho, M, 47, 55, 96, 116
Choi, T, 49, 104
Choudhery, A, 65, 159
Christou, E, 69, 175
Chu, C, 65, 160
Chu, H, 62, 67, 69, 149
Chu, J, 65, 159
Chun, H, 65, 125
Chung, D, 65, 170
Cicconetti, G, 56, 70, 104, 143
Ciccek, E, 52, 112
Ciupereca, G, 53, 117
Cook, B, 50, 106
Cook, D, 60, 141
Cook, R, 43, 43, 51, 81
Cooper, F, 47, 95
Cowles, K, 52, 119
Cox, D, 49, 104
Craige, B, 50, 62, 108, 150
Crainiceniucu, C, 58, 136
Crippa, F, 47, 74
Crowe, B, 49, 59, 103, 103
Cui, L, 59, 106, 158
Cui, Y, 54, 48, 51, 64
Curt, D, 67, 167
Curt, D, 64, 70, 158, 178
Culppepper, S, 70, 180
Cunnanan, K, 66, 164
Curtis, C, 67, 167
DAI, X, 68, 173
Dang, Q, 46, 90
D'Angelo, P, 66, 165
Daniels, M, 43, 85
Davis, C, 40, 91
Dawson, M, 71, 184
Dempster, AP, 72, 185
Deng, A, 45, 88
Deng, M, 50, 127, 128
Deng, S, 72, 185
Deng, Y, 53, 119
Dette, J, 44, 83
Devanarayan, V, 53, 118
Di, C, 48, 55, 99, 133
Diao, G, 57, 132
<table>
<thead>
<tr>
<th>Name</th>
<th>Page 1</th>
<th>Page 2</th>
<th>Page 3</th>
<th>Page 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhu, R</td>
<td>51</td>
<td>69</td>
<td>108</td>
<td>176</td>
</tr>
<tr>
<td>Zhu, Y</td>
<td>43</td>
<td>67</td>
<td>80</td>
<td>168</td>
</tr>
<tr>
<td>Zhu, Z</td>
<td>44</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zigler, C</td>
<td>45</td>
<td>87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>zou, F</td>
<td>45</td>
<td>86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zimmerman, D</td>
<td>52</td>
<td>114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zou, K</td>
<td>53</td>
<td>118</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zohar, S</td>
<td>46</td>
<td>92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zubizarreta, J</td>
<td>52</td>
<td>113</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
See you in New Brunswick New Jersey in 2018!!!

www.icsa.org