



# 2020 ICSA APPLIED STATISTICS SYMPOSIUM

Houston, Texas  
December 13-16, 2020



International Chinese Statistical Association

泛華統計協會

**International Chinese Statistical Association**

**Applied Statistics Symposium**

**2020**

**CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS**

December 13 - 16, 2020

Texas Medical Center

Houston, Texas, USA

Organized by

International Chinese Statistical Association

©2020

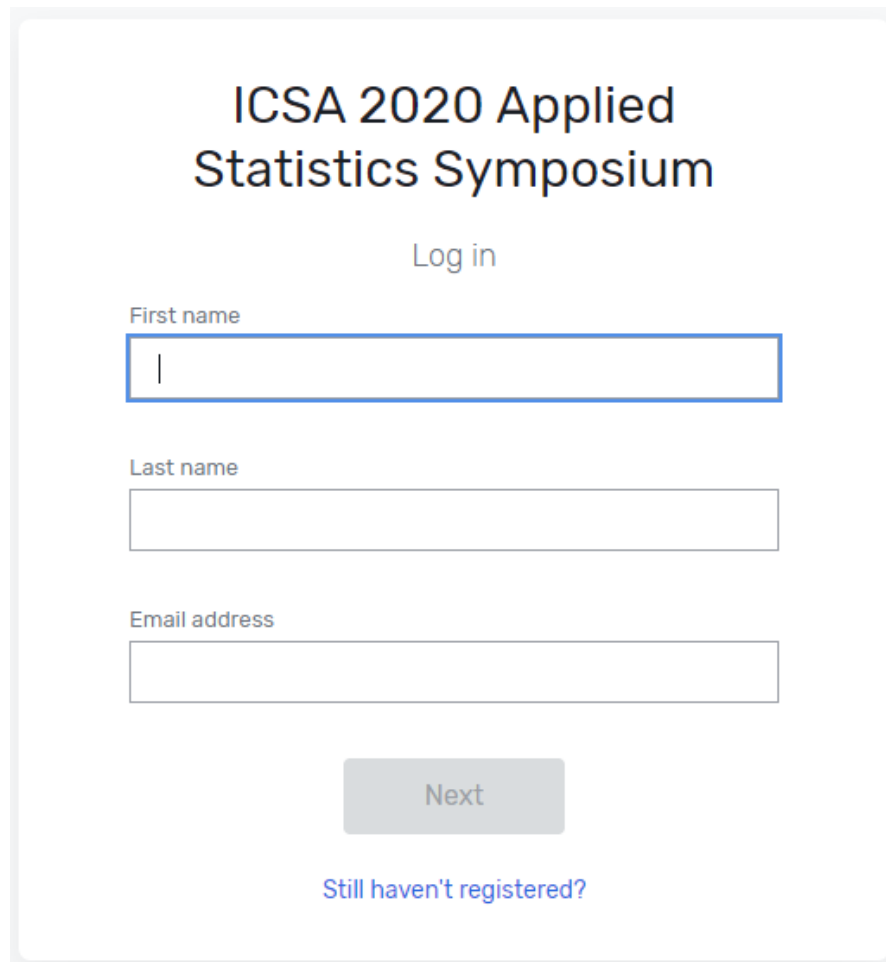
International Chinese Statistical Association

## Cvent Login Instruction

The conference is virtual! Please follow the link to login into the virtual meeting site:

<https://cvent.me/WL1xDk>

The login screen looks like this



The screenshot shows a login form for the "ICSA 2020 Applied Statistics Symposium". The form is titled "Log in" and contains three input fields: "First name", "Last name", and "Email address". The "First name" field has a vertical cursor. Below the input fields is a "Next" button, which is currently disabled (greyed out). At the bottom of the form, there is a link that says "Still haven't registered?".

- Make sure to use the email address you used in registration to login. Once you press “Next”, you will receive an email with verification code to complete the login process.
- Detail on how to participate a session, navigate inside the meeting site as well as ask for technical assistance can be found at <https://symposium2020.icsa.org/>

# Contents

Welcome . . . . .	1
Conference Information . . . . .	2
ICSA Officers and Committees . . . . .	2
Conference Committees . . . . .	5
Volunteers . . . . .	9
Sponsor Information . . . . .	10
Social Event: Talent Show Performances . . . . .	16
Program Overview . . . . .	17
Keynote Lectures . . . . .	18
Student Paper Awards) . . . . .	21
Short Courses . . . . .	22
Scientific Program . . . . .	28
Dec. 13 18:00 - 19:40 . . . . .	28
Dec. 14 9:00 - 10:00 . . . . .	29
Dec. 14 10:20 - 12:00 . . . . .	29
Dec. 14 12:20 - 13:50 . . . . .	32
Dec. 14 14:00 - 15:40 . . . . .	32
Dec. 14 16:00 - 17:40 . . . . .	35
Dec. 14 18:00 - 21:00 . . . . .	37
Dec. 15 9:00 - 10:00 . . . . .	39
Dec. 15 10:20 - 12:00 . . . . .	40
Dec. 15 12:20 - 13:50 . . . . .	42
Dec. 15 14:00 - 15:40 . . . . .	42
Dec. 15 16:00 - 17:40 . . . . .	45
Dec. 16 9:00 - 10:00 . . . . .	48
Dec. 16 10:20 - 12:00 . . . . .	48
Abstracts . . . . .	50
Session 1: Advancement of Machine Learning Methods via Tensors and High-Dimensional Tools . . . . .	50
Session 2: Latest development in latent variable models and genetics . . . . .	50
Session 3: New methods for joint analysis of survival and longitudinal data . . . . .	51
Session 4: The Data are BIG and We are PRECISE: New Statistical Methods for Precision Medicine. . . . .	52
Session 5: Decipher cell heterogeneity in high-throughput data analysis . . . . .	53
Session 6: Recent Development on Heterogeneity Analysis . . . . .	54
Session 7: Current Development in Experimental Designs and Its Applications . . . . .	54
Session 8: Keynote speech . . . . .	55
Session 9: Bayesian Methodology and Applications for Complex Biomedical Data . . . . .	55
Session 10: New Challenges in Lifetime Data Analyses . . . . .	56
Session 11: Novel Semiparametric and Machine Learning Tools in Complex Observational Studies . . . . .	57
Session 12: Statistical Inference and Modeling for High-Dimensional and Complex Data Structure . . . . .	58
Session 13: Artificial Intelligence and Causal Inference . . . . .	58
Session 14: New advances in modern statistical modeling and testing . . . . .	59
Session 15: Robust Methods in Missing Data and Causal Inference . . . . .	60
Session 16: Functional Data Analysis: Theory and Application . . . . .	61
Session 17: Recent advances in multivariate and high-dimensional statistics . . . . .	61
Session 18: Big Data Analysis: New Directions and Innovation . . . . .	62

Session 19: Recent Trends of Innovative Methodologies and Applications in Rare Disease Clinical Trials . . . . .	63
Session 20: Data Analysis and Application for High-Throughput Biotechnologies . . . . .	64
Session 21: Statistical Methods for Sports Data Analytics . . . . .	65
Session 23: Innovative statistical methods for complex survival data and the applications . . . . .	65
Session 24: Methods and applications in large and complex data . . . . .	66
Session 25: Better Evidence Syntheses in Data Science . . . . .	67
Session 26: Deciphering Multi-omics Data: Statistical Models and Computational Approaches for Biology and Health . . . . .	67
Session 27: Advanced Adaptive Enrichment Designs in Confirmative Clinical Trials . . . . .	68
Session 28: New Challenges and Opportunities in Early-Phase Oncology Trials . . . . .	69
Session 29: Novel clinical trial designs in the era of precision medicine and immunotherapy . . . . .	70
Session 30: Statistical inference and practical issues in psychiatry . . . . .	70
Session 31: Recent development in dynamic historical data borrowing: methodology and application in clinical trials . . . . .	71
Session 32: Bayesian Analysis of Complex Survey Data . . . . .	72
Session 33: Multiple phenotypes, Pleiotropy and Mendelian Randomization . . . . .	73
Session 34: New methods of clinical trial designs and analyses and sample size re-estimation . . . . .	73
Session 35: Utilization of RWE in Drug Development: Case Studies . . . . .	74
Session 36: Challenges and developments in analyzing complex data . . . . .	74
Session 37: Materials Informatics . . . . .	75
Session 38: Methodological Advances for Harmonizing Genomics Data to Enable Reproducible Biomedical Research . . . . .	76
Session 39: Statistical Method Development Motivated by Biomedical Data Challenges . . . . .	76
Session 40: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences . . . . .	77
Session 41: Statistical methods for complex human genetic data . . . . .	78
Session 42: Recent advances in statistical methods for missing data, measurement error and biased sampling . . . . .	79
Session 43: Statistical Learning Advancement for Inference with Complex Biomedical Data . . . . .	80
Session 44: New Developments in High-Dimensional Data Analysis . . . . .	81
Session 45: Empirical Likelihood Methods and Bayesian Variable Selection . . . . .	82
Session 46: Statistical Process Control and Detection of Change-Point . . . . .	82
Session 47: Novel computational techniques for analyzing large scale biostatistical data . . . . .	83
Session 48: Innovations in Statistical Machine Learning . . . . .	84
Session 49: Advances in Clinical Trial Statistics . . . . .	85
Session 50: Bayesian Statistics . . . . .	87
Session 51: Recent developments in AI and its Applications . . . . .	88
Session 52: Statistics in Genetics . . . . .	90
Session 53: Recent statistical advances in longitudinal and survival analysis . . . . .	92
Session 54: Statistical innovations in medicine and public health . . . . .	93
Session 55: Theory and Methodology for Big and Complex Data . . . . .	94
Session 56: Keynote speech . . . . .	96
Session 57: Statistical Applications of Extreme Value Theory . . . . .	96
Session 58: Variable selection with complex lifetime data . . . . .	96
Session 59: Recent advances in statistical methods for big biomedical data integration . . . . .	97
Session 60: Complex data analysis in business, economics, and industry . . . . .	98
Session 61: Bayesian Analysis of Complex and High Dimensional Data . . . . .	99
Session 62: New statistical methods for machine learning on big data . . . . .	100
Session 63: Innovative statistical methods for optimal treatment selection and clinical trial design with historical data . . . . .	101
Session 64: Statistical method advancement for analyzing omics data . . . . .	101
Session 65: Statistical learning with complex data structure . . . . .	102
Session 66: Bridging the gap between complex data and public health policies: methods and applications . . . . .	103
Session 67: Advanced Bayesian methods in Biostatistics . . . . .	103
Session 68: Leveraging Real-World Data in Comparative Effectiveness Research . . . . .	104
Session 70: Design and Statistical Issues for Pediatric Oncology Trials . . . . .	105
Session 71: Enhancing RCT using Real World Evidence . . . . .	105
Session 72: Real World Evidence for Value-Added Patient-Centric Healthcare . . . . .	106

Session 73: High-dimensional statistical learning in big-data of human genetics . . . . .	107
Session 74: Precision oncology trials: challenges and opportunities . . . . .	108
Session 75: Statistical Advancement and Challenges in Cell Therapy Development . . . . .	109
Session 76: Genetics and Genomics: Methodology and Applications . . . . .	109
Session 77: Applications of advanced statistics and artificial intelligence to genomics and precision medicine . . . . .	110
Session 78: Innovative adaptive clinical trial designs . . . . .	111
Session 79: Advanced Statistical Learning for High-dimensional Heterogeneous Data . . . . .	112
Session 80: New development in Bayesian methods and algorithms . . . . .	113
Session 81: Innovative methods for complex censored data . . . . .	114
Session 82: Student Paper Award Invited Session . . . . .	115
Session 83: Statistical Methods for design and analysis of health outcomes . . . . .	116
Session 84: Statistical Modeling for COVID-19 . . . . .	117
Session 85: Recent statistical advancements in the design of clinical trials . . . . .	118
Session 86: Challenging Statistical Issues in Oncology Studies . . . . .	118
Session 87: Recent advances in machine learning and causal inference . . . . .	119
Session 88: Recent advances in accounting for heterogeneity in complex data . . . . .	120
Session 89: Current advances in forensic statistics . . . . .	121
Session 90: Statistical and Machine Learning models on EHR and Insurance Claim databases . . . . .	122
Session 91: Utilization of Historical Control Data for Clinical Development . . . . .	123
Session 92: Statistical development for single-cell RNA-Seq data in biomedical studies . . . . .	123
Session 93: Machine Learning and Real World Data . . . . .	124
Session 94: Recent advances in statistical genomics, genetics and EHR data . . . . .	125
Session 95: Real World Evidence Study in Healthcare . . . . .	125
Session 96: Bayesian Additive Regression Tree: Theory, Computation, and Application . . . . .	126
Session 97: New Methods for Missing Data in Public Health Studies . . . . .	127
Session 98: Keynote speech . . . . .	128
Session 99: Innovative Statistical and data science methods for clinical trial studies . . . . .	128
Session 100: New methods in semiparametric inferences for analyzing real world data . . . . .	129
Session 101: Recent development in Semiparametric regression analysis . . . . .	129
Session 102: Stochastic gradient Monte Carlo for big data statistics . . . . .	130
Session 103: Statistical and AI inferences based on DNA and protein sequences . . . . .	130
Session 104: Advanced topics in causal inference . . . . .	131
Index of Authors . . . . .	132

## 2020 ICSA Applied Statistics Symposium

December 13-16, 2020, Virtual  
Organized in Houston, Texas, USA

Welcome to the 2020 International Chinese Statistical Association (ICSA) Applied Statistics Symposium!

The Year of 2020 is special. The 2020 ICSA Applied Statistics Symposium was originally scheduled from May 17-20, 2020 at Westin Galleria Houston, Houston, Texas USA. Due to COVID-19 pandemic, the symposium was moved to December 13-16, 2020 as a virtual conference. This is the 29th annual symposium for ICSA. The theme of this conference is “*Advancing Statistics for Data Intelligence*”, in recognition of a new artificial intelligence and Big Data era for statisticians with rising opportunities and challenges.

The organizing committees have been working diligently to put together a comprehensive scientific program and other activities to provide ample opportunities for discussions and exchanges of novel ideas in advancing statistics to extract knowledge from data. The symposium program contains 9 short courses, two panel discussion sessions, one special memorial session and 102 scientific sessions, including three keynote lectures, one student paper award session, and 7 poster sessions as well as an exciting event, the ICSA General Member Meeting, Awards Ceremony and Talent Show on Tuesday night. Keynote lectures are from three pioneers and distinguished statisticians from academia and industry: **Dr. Xihong Lin** (*Harvard University*), **Dr. Josh Chen** (*Sanofi Pasteur*), and **Dr. Michael I. Jordan** (*University of California at Berkeley*). The symposium highlights methodological and applied contributions of statistics, data science, mathematics, and computer sciences. It brings together the statistical and data science communities as well as scientists from related fields to present, discuss and disseminate research and best practice.

With your full support, this symposium attracts more than 600 statisticians and data scientists working in academia, government, and industry from all over the world. We hope that the symposium offers you great opportunities for learning, networking and recruiting, and that you will receive inspirations from the presented research ideas and develop new ones. In this year, we also organized 7 poster sessions with the mixer together on Monday night (6:00-9:30PM CT) and you may join the poster sessions to chat and network with your students and colleagues in the virtual room. On Tuesday night (7:30-10:00PM CT), although we cannot have a banquet together, we organized the Talent Show after the ICSA General Member Meeting and Awards Ceremony so that we can have fun together. We believe this conference will be a memorable, interesting and enjoyable experience for all of us on the cloud.

Although you could not enjoy the city of Houston with strengths in business, international trade, entertainment, culture, media, fashion, science, sports, technology, education, medicine, and research due to COVID-19, we hope that in the future you have opportunity to welcome you in Houston, home to many cultural institutions and exhibits, which attract more than 7 million visitors a year to the Museum of Fine Arts, Houston Museum of Natural Science, the Contemporary Arts Museum Houston, the Station Museum of Contemporary Art, Holocaust Museum Houston, and the Houston Zoo, and NASA Lyndon B. Johnson Space Center.

**Thank you for coming to the 2020 ICSA Applied Statistics Symposium on Cloud!**

Hulin Wu, PhD, on behalf of 2020 Applied Statistics Symposium Executive and Organizing Committees

# ICSA Officers and Committees

## ICSA 2020 EXECUTIVES AND MEMBERS OF THE COMMITTEES

### EXECUTIVES

President: Jianguo (Tony) Sun ([sunj@missouri.edu](mailto:sunj@missouri.edu))  
Past President: Heping Zhang ([heping.zhang@yale.edu](mailto:heping.zhang@yale.edu))  
President-elect: Colin Wu ([wuc@nhlbi.nih.gov](mailto:wuc@nhlbi.nih.gov))  
Executive Director: Mengling Liu  
([mengling.liu@nyulangone.org](mailto:mengling.liu@nyulangone.org))  
ICSA Treasurer: Rochelle Fu (2019-2021, [fur@ohsu.edu](mailto:fur@ohsu.edu)).

The ICSA Office Manager: Grace Ying Li,  
Email: [oicsa@icsa.org](mailto:oicsa@icsa.org), Phone: (317) 287-4261.

### BOARD of DIRECTORS

Chung-Chou H (Joyce) Chang (2018-2020, [changj@pitt.edu](mailto:changj@pitt.edu)),  
Haitao Chu (2018-2020, [chux0051@umn.edu](mailto:chux0051@umn.edu)),  
Daniel Li (2018-2020, [dli.lihe@gmail.com](mailto:dli.lihe@gmail.com)),  
Mark Chunming Li (2018-2020, [Chunming.M.Li@gmail.com](mailto:Chunming.M.Li@gmail.com)),  
Niansheng Tang (2018-2020, [nstang@ynu.edu.cn](mailto:nstang@ynu.edu.cn)).  
Yinglei Lai (2019-2021, [ylai@gwu.edu](mailto:ylai@gwu.edu)),  
Lei Shen (2019-2021, [shen\\_lei@lilly.com](mailto:shen_lei@lilly.com)),  
Yifei Sun (2019-2021, [ys3072@cumc.columbia.edu](mailto:ys3072@cumc.columbia.edu)),  
Xin Tian (2019-2021, [tianx@nhlbi.nih.gov](mailto:tianx@nhlbi.nih.gov)),  
Kelly Zou (2019-2021, [KelZouDS@gmail.com](mailto:KelZouDS@gmail.com)),  
Jason Liao (2020-2022, [jason\\_liao@merck.com](mailto:jason_liao@merck.com)),  
Bin Nan (2020-2022, [bnan@umich.edu](mailto:bnan@umich.edu)),  
Peihua Qiu (2020-2022, [pqiu@phhp.ufl.edu](mailto:pqiu@phhp.ufl.edu)),  
Jane Zhang (2020-2022, [Zhang\\_Jane@Allergan.com](mailto:Zhang_Jane@Allergan.com)),  
Yichuan Zhao (2020-2022, [yichuan@gsu.edu](mailto:yichuan@gsu.edu))

### STANDING COMMITTEES

#### Program Committee

**Zhezhen Jin** (Chair, 2020, [zj7@cumc.columbia.edu](mailto:zj7@cumc.columbia.edu)); Guoqing Diao (2018-2020, [gdiao@gmu.edu](mailto:gdiao@gmu.edu)), Xiwu Lin (2018-2020, [xlin38@ITS.JNJ.com](mailto:xlin38@ITS.JNJ.com)), Wenbin Lu (2018-2020, [wlu4@ncsu.edu](mailto:wlu4@ncsu.edu)), Liuquan Sun (2018-2020, [slq@amt.ac.cn](mailto:slq@amt.ac.cn)), Kai Yu (2018-2020, [yuka@mail.nih.gov](mailto:yuka@mail.nih.gov)), Alan Y Chiang (2019-2021, [achiang@celgene.com](mailto:achiang@celgene.com)), Bin Nan (2019-2021, [nanb@uci.edu](mailto:nanb@uci.edu)), Ji Zhu (2019-2021, [jjzhu@umich.edu](mailto:jjzhu@umich.edu)), Qingning Zhou (2020-2022, [qzhou8@ucc.edu](mailto:qzhou8@ucc.edu)), Liang Zhu (2020-2022, [Liang.Zhu@uth.tmc.edu](mailto:Liang.Zhu@uth.tmc.edu)), Hulin Wu, (2020-2022, [Hulin.Wu@uth.tmc.edu](mailto:Hulin.Wu@uth.tmc.edu)), Jie Chen (2020-2022, [jiechen0713@gmail.com](mailto:jiechen0713@gmail.com))

#### Finance Committee

**Rochelle Fu** (Chair, 2019-2021, [fur@ohsu.edu](mailto:fur@ohsu.edu)); Hongliang Shi ([hongliangshi15@gmail.com](mailto:hongliangshi15@gmail.com)), Rui Feng ([ruifeng@penntestmed.upenn.edu](mailto:ruifeng@penntestmed.upenn.edu)), Shu Yang ([syang24@ncsu.edu](mailto:syang24@ncsu.edu)).



# ICSA Officers and Committees

## Nomination for Election Committee

**Joan Hu** (Chair, 2020, [joan\\_hu@sfu.ca](mailto:joan_hu@sfu.ca)); Chin-Tsang Chiang (2018-2020, [chiangct@math.ntu.edu.tw](mailto:chiangct@math.ntu.edu.tw)), Jeen Liu (2018-2020, [Liu\\_jeen@allergan.com](mailto:Liu_jeen@allergan.com)), Ming Tan (2018-2020, [Ming.Tan@georgetown.edu](mailto:Ming.Tan@georgetown.edu)), Xin Tian (2018-2020, [tianx@nhlbi.nih.gov](mailto:tianx@nhlbi.nih.gov)), Bo Huang (2019-2021, [Bo.Huang@pfizer.com](mailto:Bo.Huang@pfizer.com)), Chunling Liu (2019-2021, [catherine.chunling.liu@polyu.edu.hk](mailto:catherine.chunling.liu@polyu.edu.hk)), Yiyuan She (2019-2021, [yshe@stat.fsu.edu](mailto:yshe@stat.fsu.edu)), Jiayang Sun (2019-2021, [jsun@case.edu](mailto:jsun@case.edu)), Hailong Cheng (2020-2022, [hailong.cheng@sunovion.com](mailto:hailong.cheng@sunovion.com)), Bin Zhang (2020-2022, [Bin.Zhang@cchmc.org](mailto:Bin.Zhang@cchmc.org))

## Publication Committee

**Yichuan Zhao** (2020, [yichuan@gsu.edu](mailto:yichuan@gsu.edu)), Yi Huang (Editor of Bulletin, [yihuang@umbc.edu](mailto:yihuang@umbc.edu)), Mei-Cheng Wang (Co-Editor of SIB, [mcwang@jhu.edu](mailto:mcwang@jhu.edu)), Hongzhe Li (Co-Editor of SIB, [hongzhe@mail.med.upenn.edu](mailto: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), Jiahua Chen (Editor of ICSA book series), Din Chen (Co-Editor of ICSA book series).

## Membership Committee

**Bo Fu** (Chair, 2020, [bo.fu@abbvie.com](mailto:bo.fu@abbvie.com)); Liuquan Sun (Co-Chair, 2020, [slq@amt.ac.cn](mailto:slq@amt.ac.cn)), Mark Li (2018-2020, [Chunming.M.Li@pfizer.com](mailto:Chunming.M.Li@pfizer.com)), Chenguang Wang (2018-2020, [cwang68@jhmi.edu](mailto:cwang68@jhmi.edu)), Yaping Wang (2018-2020, [yapping.wang@fda.hhs.gov](mailto:yapping.wang@fda.hhs.gov)), Niansheng Tang (2019-2021, [nstang@ymu.edu.cn](mailto:nstang@ymu.edu.cn)), Lei Shen (2019-2021, [shen\\_lei@lilly.com](mailto:shen_lei@lilly.com)), Shuwei Li (2020-2022, [lishuwstat@163.com](mailto:lishuwstat@163.com))

## Special Lecture Committee

**Jianqing Fan** (Chair, 2020, [jianqing.fan@outlook.com](mailto:jianqing.fan@outlook.com)); Ming Yuan (2018-2020, [ming.mingyuan@gmail.com](mailto:ming.mingyuan@gmail.com)), Zhezhen Jin (2018-2020, [zj7@cumc.columbia.edu](mailto:zj7@cumc.columbia.edu)), Gang Li of UCLA, (2018-2020, [vli@g.ucla.edu](mailto:vli@g.ucla.edu)), Haiqun Lin (2019-2021, [haiqun.lin@yale.edu](mailto:haiqun.lin@yale.edu)), Huazhen Lin (2019-2021, [linhz@swufe.edu.cn](mailto:linhz@swufe.edu.cn)).

## Awards Committee

**Xiangrong Yin** (Chair, 2020, [yinxiangrong@uky.edu](mailto:yinxiangrong@uky.edu)); Frank Guanghan Liu (2018-2020, [guanghan\\_frank\\_liu@merck.com](mailto:guanghan_frank_liu@merck.com)), Mei-Cheng Wang (2018-2020, [mcwang@jhu.edu](mailto:mcwang@jhu.edu)), Song Yang (2018-2020, [yangso@nhlbi.nih.gov](mailto:yangso@nhlbi.nih.gov)), Jie Chen (2019-2021, [jiechen0713@gmail.com](mailto:jiechen0713@gmail.com)), Ning Hao (2019-2021, [nhao@math.arizona.edu](mailto:nhao@math.arizona.edu)), Judy Wang (2019-2021, [judywang@gwu.edu](mailto:judywang@gwu.edu)), Hongyuan Cao (2020-2022, [hcao@fsu.edu](mailto:hcao@fsu.edu)), Xiaogang Su (2020-2022, [xsu@utep.edu](mailto:xsu@utep.edu)), Xiaofeng Shao (2020-2022, [xshao@illinois.edu](mailto:xshao@illinois.edu)), Yichao Wu (2020-2022, [yichaowu@uic.edu](mailto:yichaowu@uic.edu)).

## Financial Advisory Committee

**Rui Feng** (Chair, [ruifeng@upenn.edu](mailto:ruifeng@upenn.edu)); Hongliang Shi ([hongliangshi15@gmail.com](mailto:hongliangshi15@gmail.com)), Nianjun Liu, ([liunian@indiana.edu](mailto:liunian@indiana.edu)), Xiangqin Cui ([xiangqin.cui@emory.edu](mailto:xiangqin.cui@emory.edu)), Xiangqin Cui ([xiangqin.cui@emory.edu](mailto:xiangqin.cui@emory.edu)), Fang Chen ([FangK.Chen@sas.com](mailto:FangK.Chen@sas.com)), Rochelle Fu ([fur@ohsu.edu](mailto:fur@ohsu.edu)).

## IT Committee

**Chengsheng Jiang** (Chair, 2020, [chengsheng.jiang@gmail.com](mailto:chengsheng.jiang@gmail.com)).

## Archive Committee

**Chung-Chou H (Joyce) Chang** (Chair, 2020, [changj@pitt.edu](mailto:changj@pitt.edu)); Xin Tian (2019-2021, [tianx@nhlbi.nih.gov](mailto:tianx@nhlbi.nih.gov))

## Lingzi Lu Award Committee (ASA/ICSA)

**Chan, Ivan** (Chair, 2019-2021, [ivan.chan@abbvie.com](mailto:ivan.chan@abbvie.com)); Jichun Xie (2018-2020, Duke University, [jichun.xie@duke.edu](mailto:jichun.xie@duke.edu)), Shelly Hurwitz (2017-2022, [hurwitz@hms.harvard.edu](mailto:hurwitz@hms.harvard.edu)), Laura J Meyerson (2020-2022, [laurameyerson@msn.com](mailto:laurameyerson@msn.com))

## ICSA Representative to JSM Program Committee

Xuming He (2020, [xmhe@umich.edu](mailto:xmhe@umich.edu)).

## AD HOC COMMITTEES

### 2020 Applied Statistics Symposium

Hulin Wu ([Hulin.Wu@uth.tmc.edu](mailto:Hulin.Wu@uth.tmc.edu)), Chair of Executive Committee.

### 2020 ICSA China Conference

Ying Zhang ([ying.zhang@unmc.edu](mailto:ying.zhang@unmc.edu)), Chair of the Program Committee, and Hui Zhao ([hzhao@zuel.edu.cn](mailto:hzhao@zuel.edu.cn)), Co-Chair of the Program Committee.

### 2019 JSM Local Committee

Yong Chen (Chair)

## CHAPTERS

### ICSA-Canada Chapter

Liqun Wang (Chair, [Liqun.Wang@umanitoba.ca](mailto:Liqun.Wang@umanitoba.ca))

### ICSA-Midwest Chapter

Li Wang (Chair, [li.wang1@abbvie.com](mailto:li.wang1@abbvie.com))

### ICSA-Taiwan Chapter

Chao A. Hsiung (Chair, [hsiung@nhri.org.tw](mailto:hsiung@nhri.org.tw))

## Executive Committee:

- Executive Committee Chair: Hulin Wu, University of Texas Health Science Center at Houston
- Scientific Program Co-Chair: Momiao Xiong , University of Texas Health Science Center at Houston
- Scientific Program Co-Chair: Jianhua Huang , Texas A&M University
- Poster Session Committee Chair: Xi Luo, University of Texas Health Science Center at Houston
- Program Book and Website Committee Co-Chair: Yunxin Fu, University of Texas Health Science Center at Houston
- Program Book and Website Committee Co-Chair: Ashraf Yaseen, University of Texas Health Science Center at Houston
- Local Committee Chair: Hongyu Miao, University of Texas Health Science Center at Houston
- Treasurer: Dejian Lai, University of Texas Health Science Center at Houston
- Student Paper Competition Committee Co-Chair: Ruosha Li, University of Texas Health Science Center at Houston
- Student Paper Competition Committee Co-Chair: Jing Ning, University of Texas MD Anderson Cancer Center
- Short Course Committee Chair: Wenyi Wang, University of Texas MD Anderson Cancer Center
- Fund Raising Committee Chair: Rui (Sammi) Tang, Servier Pharmaceuticals
- Talent Show Committee Chair: Kelly H. Zou, Viatrix
- Strategic Advisor: Zhezhen Jin, Columbia University
- Strategic Advisor: Wenbin Lu, North Carolina State University
- Strategic Advisor: Lanju Zhang, Abbvie Inc
- Conference Secretary: Gen Zhu (Student Volunteer), University of Texas Health Science Center at Houston

## Scientific Program Committee:

- Co-Chair: Momiao Xiong, University of Texas Health Science Center at Houston
- Co-Chair: Jianhua Huang, Texas A&M University
- Abidemi Adeniji, resTORbio Inc.
- Baojiang Chen, University of Texas Health Science Center at Houston
- Colin O. Wu, NIH/NHLBI
- Faming Liang, Purdue University
- Fengyu Wang, Tianjin University
- Henghsiu Tsai, Academia Sinica
- Hongjian Zhu, University of Texas Health Science Center at Houston
- Hong-Wen Deng, Tulane University
- Hua Tang, Stanford University
- Jack Lee, University of Texas MD Anderson Cancer Center

- Lei Wang, The Lotus Group
- Liang Fang, MyoKardia Inc.
- Liang Li, University of Texas MD Anderson Cancer Center
- Liang Zhu, University of Texas Health Science Center at Houston
- Linbo Wang, University of Toronto
- Liping Zhu, Renmin University of China
- Lixing Zhu, Hong Kong Baptist University
- Mei-Ling Ting Lee, University of Maryland
- Min-Qian Liu, Nankai University
- Peng Wei, University of Texas MD Anderson Cancer Center
- Ruixiao Lu, Genomic Health
- Ruzong Fan, Georgetown University
- Ryan Sun, University of Texas MD Anderson Cancer Center
- Suojin Wang, Texas A&M University
- Tony Jiang, Amgen Inc.
- Wanyang Dai, Nanjing University
- Wenjiang Fu, University of Houston
- Wuji Li, Institute of Basic Medical Science Research in China
- Xi Luo, University of Texas Health Science Center at Houston
- Xiaofeng Zhu, Case Western Reserve University
- Xiaohua Zhang, University of Macau
- Xingqiu Zhao, The Hong Kong Polytechnic University
- Xinsheng Zhang, Fudan University
- Xinyuan Song, The Chinese University of Hong Kong
- Ye Shen, University of Georgia
- Yi-Ching Yao, Academia Sinica
- Yihua (Mary) Zhao, Boehringer Ingelheim Pharmaceuticals, Inc.
- Ying Lu, Stanford University
- Ying Nian Wu, University of California at Los Angeles
- Ying Yuan, University of Texas MD Anderson Cancer Center
- Yu shen, University of Texas MD Anderson Cancer Center
- Yunxiao Chen, London School of Economics and Political Science
- Zhaoyang Teng, Servier Pharmaceuticals

### **Poster Session Committee:**

- Chair: Xi Luo , University of Texas Health Science Center at Houston
- Cici Bauer, University of Texas Health Science Center at Houston
- Jian Kang, University of Michigan
- Jose-Miguel Yamal, University of Texas Health Science Center at Houston
- Michele Guindani, University of California Irvine
- Yong Chen, University of Pennsylvania

### **Student Paper Competition Committee:**

- Co-Chair: Ruosha Li, University of Texas Health Science Center at Houston
- Co-Chair: Jing Ning, University of Texas MD Anderson Cancer Center
- Kehui Chen, University of Pittsburgh
- Yisheng Li, University of Texas MD Anderson Cancer Center
- Ruitao Lin, University of Texas MD Anderson Cancer Center
- Suyu Liu, University of Texas MD Anderson Cancer Center
- Yang Ni, Texas A&M University
- Haitao Pan, St. Jude Children's Research Hospital
- Xinlei (sherry) Wang, Southern Methodist University
- Qi Zheng, University of Louisville

### **Short Course Committee:**

- Chair: Wenyi Wang, University of Texas MD Anderson Cancer Center
- Bin Zhu, NIH/NCI
- Jack Lee, University of Texas MD Anderson Cancer Center
- Jason Liao, Merck
- Meng Li, Rice University
- Nan Shao, Boehringer Ingelheim
- Wenjiang Fu, University of Houston
- Yaohua Zhang, Vertex

### **Financial/Business Committee:**

- Chair/Treasurer: Dejian Lai, University of Texas Health Science Center at Houston
- Yisheng Li, University of Texas MD Anderson Cancer Center

### **Fund Raising Committee:**

- Chair: Rui (Sammi) Tang, Servier Pharmaceuticals
- Dejian Lai, University of Texas Health Science Center at Houston
- Hui Yang, Amgen
- Honghong Zhou, Moderna
- Ruixiao Lu, Exact Sciences
- Xiaohua Sheng, ClinChoice

### **Local Committee:**

- Chair: Hongyu Miao, University of Texas Health Science Center at Houston
- Cici Bauer, University of Texas Health Science Center at Houston
- Hongjian Zhu, University of Texas Health Science Center at Houston
- Peng Wei, University of Texas MD Anderson Cancer Center
- Ryan Sun, University of Texas MD Anderson Cancer Center
- Wei Zhang, University of Texas Health Science Center at Houston

## **Program Book and Website Committee**

- Co-Chair: Yunxin Fu, University of Texas Health Science Center at Houston
- Co-Chair: Ashraf Yaseen, University of Texas Health Science Center at Houston
- ICSA IT: Chengsheng Jiang, University of Maryland
- Web master: Gen Zhu, University of Texas Health Science Center at Houston

## **Talent Show Committee**

- Chair: Kelly H. Zou, Viatrix
- Haoda Fu, Eli Lilly and Company
- Chengsheng Jiang, University of Maryland
- Aiyi Liu, NIH
- Mina Riad, Viatrix
- Jie Tang, Lotus Clinical Research
- Qiuyi Wu, University of Rochester
- Gen Zhu, University of Texas Health Science Center at Houston

<b>Name</b>	<b>School</b>
Jingxiao Chen	UTHealth
Medina Colic	MD Anderson
Biai Dominique Elmir Digbeu	UTHealth
Han Feng	UTHealth
Yuxuan Gu	UTHealth
Jacky Kuo	UTHealth
Wenhao Li	MD Anderson
Yuan Li	UTHealth
Zhouxuan Li	UTHealth
Yiyun Lin	MD Anderson
Maya Mridula Magesh Kumar	UTHealth
Qian Qian	UTHealth
Sarah Valencia	UTHealth
Feng Wang	UTHealth
Jingyan Wang	UTHealth
Qin Wang	UTHealth
Xueying Wang	UTHealth
Heping Wang	UTHealth
Gabriella Wu	Texas A&M
Juntao Yan	UTHealth
Chao Yang	MD Anderson
Duo Yu	UTHealth
Chenguang Zhang	UTHealth
Wen Zhang	UTHealth
Kehe Zhang	UTHealth
Chengxue Zhong	UTHealth
Gen Zhu	UTHealth
Ginny Zhu	UTHealth
Jinhao Zou	MD Anderson

## Sponsors and Career Services

The 2020 ICSA Applied Statistics Symposium Program Committees gratefully acknowledge the generous support of our sponsors.

### Gold sponsors



GILEAD

abbvie



ClinChoice



Boehringer  
Ingelheim

### Silver sponsors



Daiichi-Sankyo



**MERCK**  
INVENTING FOR LIFE

### Additional sponsors



Edwards

**NEKTAR**<sup>®</sup>



# abbvie

PEOPLE. PASSION. POSSIBILITIES.®

AbbVie is a global, research-driven biopharmaceutical company committed to developing innovative advanced therapies for some of the world's most complex and critical conditions. The company's mission is to use its expertise, dedicated people and unique approach to innovation to markedly improve treatments across four primary therapeutic areas: immunology, oncology, virology and neuroscience. In more than 75 countries, AbbVie employees are working every day to advance health solutions for people around the world. For more information about AbbVie, please visit us at [www.abbvie.com](http://www.abbvie.com). Follow @abbvie on Twitter, Facebook or LinkedIn.





## Biostatistics

- Randomization
- IDMC/DSMB
- SAP
- Protocol/CSR
- ISS/ISE

## Data Management

- EDC Database Development
- Data Review and Query Management
- Dictionary Coding

## Statistical Programming

- SDTM/ADaM
- Table/Figure/Listing
- Legacy Data Conversion
- SDRG/ADRG
- eSubmission Ready

## Drug Safety

- Case Entry
- Narrative Writing
- SAE Reconciliation
- Argus Safety Database
- Global Case Submission

**Proud To Be ICSA Sponsor  
Three Years In A Row!**

Value through  
innovation

*Improving the health  
of humans and animals  
– Our goal.*

Family-owned since 1885, Boehringer Ingelheim is one of the leading pharmaceutical companies worldwide. More than 51,000 employees create value through innovation in the business areas Human Pharma, Animal Health and Biopharmaceutical Contract Manufacturing. In our role as our patients' partner we concentrate on researching and developing innovative medicines and therapies that can improve and extend patients' lives.

## Sponsors and Career Services



For more than 125 years, Merck, known as MSD outside of the United States and Canada, has been inventing for life, bringing forward medicines and vaccines for many of the world's most challenging diseases in pursuit of our mission to save and improve lives. We demonstrate our commitment to patients and population health by increasing access to health care through far-reaching policies, programs and partnerships. Today, Merck continues to be at the forefront of research to prevent and treat diseases that threaten people and animals — including cancer, infectious diseases such as HIV and Ebola, and emerging animal diseases — as we aspire to be the premier research-intensive biopharmaceutical company in the world. For more information, visit [www.merck.com](http://www.merck.com) and connect with us on Twitter, Facebook, Instagram, YouTube and LinkedIn.



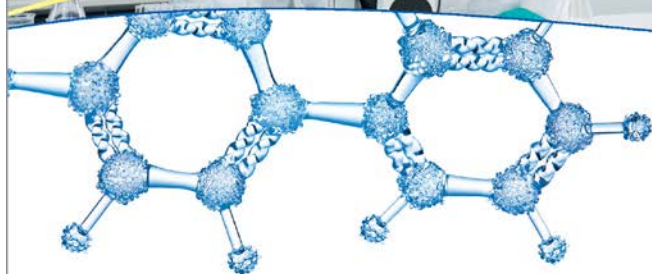
[www.tlgcareers.com](http://www.tlgcareers.com)

The Lotus Group LLC is a leader in pharmaceutical industry recruitment. We are honored to be recognized as one of the top 50 fastest-growing women-owned businesses in 2020 by the Women Presidents Organization (sponsored by American Express). If you are a talented statistician, statistical programmer or data management professional, we have a job for you!

Our recruiting team members are located throughout the country including California, New Jersey, and Massachusetts and beyond. We pride ourselves on providing excellent service and building solid relationships. To our clients, we provide top level candidates. And to our candidates, your career is in solid hands given our consultative approach regarding professional development, interview preparation and compensation guidance. Many of our knowledgeable Lotus members have 10-20+ years of experience in the pharmaceutical industry. Take a journey into success on the Lotus Group bridge - connecting candidates and companies. We look forward to meeting you!

# Sponsors and Career Services

Passion for Innovation.  
Compassion for Patients.™



 Daiichi-Sankyo

With over 100 years of scientific expertise, Daiichi Sankyo Group is dedicated to the creation and supply of innovative pharmaceutical products to address diversified, unmet medical needs of patients in both mature and emerging markets.

Under the Group's 2025 Vision to become a "Global Pharma Innovator with a Competitive Advantage in Oncology," Daiichi Sankyo research and development is primarily focused on bringing forth novel therapies in oncology, as well as other research areas centered around rare diseases and immune disorders. Learn more at [www.dsi.com](http://www.dsi.com).

For rewarding full-time employment and internship opportunities in biostatistics, programming, data management and more, visit <https://careers.dsi.com>.

Join us on Tuesday evening for

## ICSA 2020 “Brilliance!” Talent Show Performances

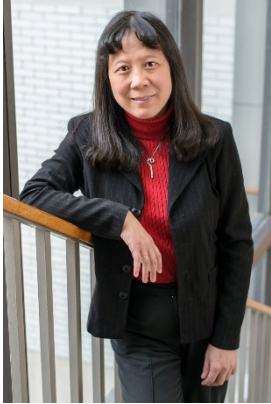
Category	Performance
Group – Platinum & Grand Prize	River Flows in You (Instruments)
Group - Gold	China at Heart Over a Jug of Wine (Singing; Photography)
Group - Silver	The Blue Sky Captains (Variety)
Group - Bronze	Tomorrow Will Be Better (Chorus)
Solo - Platinum	Colorful Clouds Chasing the Moon (Instrument)
Solo – Gold	Memory (Singing)
Solo – Silver	Flamenco Fusion (Sports)
Solo – Bronze	The Wind Shakes the Wheat (Singing)
Group	Jasmin Flower (Chorus)
Group	Life Can Be Good (Pets)
Solo	Look Back to the Beauty Around You (Singing; Photograph)
Solo	Looking Forward (Singing)
Group	Noodles (Cooking)
Group	Sentimental Values (Variety)
Group	The Wind Shakes the Wheat (Singing; Sign Language)

# Program Overview

## Activity/Event Schedules (All times are US Central Time)

Time	Location	Activity	Host
<b>12/13: Sunday</b> Virtual Front Desk Open from 8:00AM-8:00PM			
8:30-12:30PM	Breakout rooms	Morning short courses	Short Course Committee
12:30-1:30PM	Lunch break		
1:30-5:30PM	Breakout rooms	Afternoon short courses	Short Course Committee
8:30-5:30PM	Breakout rooms	Full-day courses	Short Course Committee
6:00-7:40PM	Breakout rooms	Parallel Sessions	Session Chairs
<b>12/14: Monday</b> Virtual Front Desk Open from 8:00AM-10:00PM			
8:30-9:00AM	Virtual Main Room	Welcome and opening ceremony	Hulin Wu
9:00-10:00AM		Keynote Talk I: Xihong Lin	Jianguo (Tony) Sun
10:00-10:20AM	Coffee Break		
10:20-12:00PM	Breakout rooms	Parallel Sessions	Session Chairs
12:00-12:20PM	Lunch break:		
12:20-1:50PM	Virtual Main Room	Panel Discussion Sessions: Leadership	Session Chairs
2:00-3:40PM	Breakout rooms	Parallel Sessions	Session Chairs
3:40-4:00PM	Coffee Break		
4:00-5:40PM	Breakout rooms	Parallel Sessions	Session Chairs
6:00-9:30PM	Virtual Poster & Mixer Room	Poster Sessions & Mixer	Poster Session Chairs
<b>12/15: Tuesday</b> Virtual Front Desk Open from 8:30AM-7:30PM			
9:00-10:00AM	Virtual Main Room	Keynote Talk II: Josh Chen	Gang Li
10:00-10:20AM	Coffee Break		
10:20-12:00PM	Breakout rooms	Parallel Sessions	Session Chairs
12:00-12:20PM	Lunch break		
12:20-1:50PM	Virtual Main Room	Panel Discussion Sessions: RWE Panel	Session Chairs
2:00-3:40PM	Breakout rooms	Parallel Sessions	Session Chairs
3:40-4:00PM	Coffee Break		
4:00-5:40PM	Breakout rooms	Parallel Sessions	Session Chairs
7:30-10:00PM	ICSA Zoom Room	ICSA 2020 General Member Meeting, Awards Ceremony and Talent Show	Mengling Liu Hulin Wu Kelly Zou
<b>12/16: Wednesday</b> Virtual Front Desk Open from 8:30AM-10:30AM			
9:00-10:00AM	Virtual Main Room	Keynote Talk III: Michael Jordan	Hulin Wu
10:00-10:20AM	Coffee Break		
10:20-12:00PM	Breakout rooms	Parallel Sessions	Session Chairs
12:00PM	Adjournment		

## Keynote Lectures



**Xihong Lin, Ph.D.**, Professor and former Chair of the Department of Biostatistics, Coordinating Director of the Program in Quantitative Genomics at the Harvard T. H. Chan School of Public Health, and Professor of the Department of Statistics at the Faculty of Arts and Sciences of Harvard University, and Associate Member of the Broad Institute of Harvard and MIT. Dr. Lin is an elected member of the National Academy of Medicine. She received the 2002 Mortimer Spiegelman Award from the American Public Health Association, and the 2006 Committee of Presidents of Statistical Societies (COPSS) Presidents' Award and the 2017 COPSS FN David Award. She is an elected fellow of American Statistical Association (ASA),

Institute of Mathematical Statistics, and International Statistical Institute. Dr. Lin is the former Chair of the COPSS (2010-2012) and a former member of the Committee of Applied and Theoretical Statistics (CATS) of the National Academy of Science. She co-launched the new Section of Statistical Genetics and Genomics of the ASA and served as a former section chair. She is the former Coordinating Editor of Biometrics and the founding co-editor of Statistics in Biosciences. She has served on a large number of committees of many statistical societies, and numerous NIH and NSF review panels. Dr. Lin's research interests lie in development and application of scalable statistical and computational methods for analysis of massive data from genome, exposome and phenome, and scalable statistical inference and learning for big health and genomic data. Her theoretical and computational statistical research includes statistical methods for testing a large number of complex hypotheses, statistical inference for large covariance matrices, prediction models using high-dimensional data, and cloud-based statistical computing.

**Time: December 14 (Monday): 9:00-10:00AM (Central Time)**

**Host:** Jianguo Sun, Ph.D., ICSA President and Professor, Department of Statistics, University of Missouri

**Title:** Learning from COVID-19 Data in Wuhan, USA and the World on Transmission, Health Outcomes and Interventions

**Abstract:** COVID-19 is an emerging respiratory infectious disease that has become a pandemic. In this talk, I will first provide a historical overview of the epidemic in Wuhan. I will provide the analysis results of 32,000 lab-confirmed COVID-19 cases in Wuhan to estimate transmission rates, the multi-faceted public health intervention effects that helped Wuhan control the COVID-19 outbreak, and epidemiological characteristics of the cases. I will present the results using the transmission dynamic model that show two features of the COVID-19 epidemic: high transmissibility and high covertness, and a high proportion of undetected cases, including asymptomatic and mildly symptomatic cases, and the chances of resurgence in different scenarios. I will next present the epidemic models to estimate the transmission rates in USA and other countries and intervention effects, as well as the prevalence and the total number of infections. I will present methods and analysis results of >500,000 participants of the HowWeFeel project on symptoms and health conditions in US, and discuss the factors associated with who have been tested in US and the factors associated with positive PRC tests/COVID-19 infection. I will provide several takeaways learned from the pandemic and discuss priorities.





**Josh Chen, Ph.D.**, Head of Global Biostatistical Sciences at Sanofi Pasteur. His team provides quantitative leadership through life cycle of vaccines, including discovery research, toxicology, CMC, biomarker strategy, translational sciences, clinical development strategy, study design, medical affairs, and value generation for payers. Josh promotes value and impact of quantitative scientists in the biopharmaceutical industry, and drives research and application of innovative statistical methods. His research interest includes clinical trial group sequential methods, adaptive designs and multiregional clinical trials (MRCTs). Josh's research collaboration has led to

publication of a book on the best practices for simultaneous global development and 50+ papers in peer-reviewed journals. Josh was a co-lead of the across-industry MRCT Consistency Working Group. He is a life time member of ICSA, was a member of the ICSA Board of Directors, and served as a program co-chair for the 2008 ICSA Applied Symposium. Josh received his PhD in Statistics from the University of Wisconsin-Madison, and Master and Bachelor degrees in Probability and Statistics from Peking University. He is a Fellow of the American Statistical Association.

**Time: December 15 (Tuesday): 9:00-10:00AM (Central Time)**

**Host:** Gang Li, Ph.D., Director, Statistics and Decision Science, Janssen Research & Development

**Title:** Use of Real World Healthcare Data to Accelerate Vaccine Development in the Post COVID Era

**Abstract:** Human vaccine research and development is a lengthy, risky and expensive process which typically takes 10-15 years from discovery to approval. Lessons learned from the current collaborative efforts to develop safe and effective COVID-19 vaccines within 12-18 months support the aspiration that it is possible to accelerate vaccine development using innovative approaches. Before the COVID pandemic, there had been strong interest in the potential use of real-world evidence for regulatory purposes. The COVID pandemic will further catalyzes digital transformation and advancement of information technology infrastructure and as a result, vast increase in high quality real world data pertaining to patient health and healthcare delivery. In this talk, we will advocate use of real world data from healthcare systems, including electronic health records (EHRs), medical claims and billing data, and patient registries, to generate fit-for-purpose real world evidence in support of the safety and effectiveness of an experimental vaccine for regulatory decisions.



**Michael I. Jordan, Ph.D.**, Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley. His research interests bridge the computational, statistical, cognitive and biological sciences, and have focused in recent years on Bayesian nonparametric analysis, probabilistic graphical models, spectral methods, kernel machines and applications to problems in distributed computing systems, natural language processing, signal processing and statistical genetics. Prof. Jordan is a member of the National Academy of Sciences, a member of the National Academy of Engineering and a member of the American Academy of Arts and Sciences. He is a Fellow of the American Association for the Advancement of Science. He has been named a Neyman Lecturer and a Medallion Lecturer by the Institute of Mathematical Statistics. He received the IJCAI Research Excellence Award in 2016, the David E. Rumelhart Prize in 2015 and the ACM/AAAI Allen Newell Award in 2009. He is a Fellow of the AAAI, ACM, ASA, CSS, IEEE, IMS, ISBA and SIAM.

**Time: December 16 (Wednesday): 9:00-10:00AM (Central Time)**

**Host:** Hulin Wu, Ph.D., 2020 ICSA Applied Statistics Symposium Organizing Committee Chair,

The Betty Wheless Trotter Professor & Chair, Department of Biostatistics & Data Science,

University of Texas Health Science Center at Houston

**Talk title:** Towards a Blend of Statistics and Microeconomics

**Abstract:** Statistical decisions are often given meaning in the context of other decisions, particularly when there are scarce resources to be shared. Managing such sharing is one of the classical goals of microeconomics, and it is given new relevance in the modern setting of large, human-focused datasets, and in data-analytic contexts such as classifiers and recommendation systems. I'll discuss several recent projects that aim to explore this interface, including the study of exploration-exploitation trade-offs for bandits that compete over a scarce resource, notions of local optimality in nonconvex-nonconcave minimax optimization and how such notions relate to stochastic gradient methods, the use of Langevin-based algorithms for Thompson sampling, and multi-agent learning based on online gradient descent.

## ***ICSA Applied Statistics Symposium Student Paper Awards***

- Xinyue Qi, University of Texas Health Science Center at Houston  
Title: Bayesian Meta-analysis of Censored Rare Events with Stochastic Coarsening  
*See session 82 in the program*
- Xinjun Wang, University of Pittsburgh  
Title: BREM-SC: A Bayesian Random Effects Mixture Model for Joint Clustering Single Cell Multi-omics Data  
*See session 82 in the program*
- Zhengjia Wang, Rice University  
Title: Functional Group Bridge Regression with Application to iEEG Data  
*See session 82 in the program*
- Yizhen Xu, Johns Hopkins Bloomberg School of Public Health  
Title: Inference for BART with Multinomial Outcomes  
*See session 82 in the program*
- Huijuan Zhou, Renmin University of China and Texas A&M University  
Title: Covariate Adaptive Family-wise Error Rate Control for Genome-Wide Association Studies  
*See session 82 in the program*

## ***Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper Award***

- Peng Jin, New York University Grossman School of Medicine  
Title: Generalized Mean Residual Life Models for Case-Cohort and Nested Case-Control Studies  
*See session 40 in the program*

**SC01: Multivariate meta-analysis methods****Length:** Half-day**Instructors:** Dr. Haitao Chu (University of Minnesota Twin Cities); Dr. Yong Chen (University of Pennsylvania)

**Outline/Description:** Comparative effectiveness research aims to inform health care decisions concerning the benefits and risks of different prevention strategies, diagnostic instruments and treatment options. A meta-analysis is a statistical method that combines results of multiple independent studies to improve statistical power and to reduce certain biases compared to individual studies. Meta-analysis also has the capacity to contrast results from different studies and identify patterns and sources of disagreement among those results. The increasing number of prevention strategies, assessment instruments and treatment options for a given disease condition have generated a need to simultaneously compare multiple options in clinical practice using rigorous multivariate meta-analysis methods. This short course, co-taught by Drs. Chu and Chen who have collaborated on this topic for more than a decade, will focus on most recent developments for multivariate meta-analysis methods. This short course will offer a comprehensive overview of new approaches, modeling, and applications on multivariate meta-analysis. Specifically, the instructors will discuss the contrast-based and arm-based network meta-analysis methods for multiple treatment comparisons; network meta-analysis methods for multiple diagnostic tests; and multivariate meta-analysis methods estimating complier average causal effect in randomized clinical trials with noncompliance. Case studies will be used to illustrate the principles and statistical methods introduced in this course. This application oriented short course should be of interest to researchers who would apply up-to-date multivariate meta-analysis methods. We anticipate that it will be well-received by an interdisciplinary scientific community, and play an important role in improving the rigor and broadening the applications of multivariate meta-analysis.

**About the Instructors:** Dr. Chu is Professor of Biostatistics at University of Minnesota Twin Cities. He is an ASA Fellow and elected member of the Society for Research Synthesis Methodology since 2016. Dr. Chu's research lies at the intersection of biostatistics and epidemiology, with a recent focus on multivariate research synthesis methods. Dr. Chu has published over 170 peer-reviewed articles with over 10,000 Google Scholar citations. Specifically, Dr. Chu has published over 50 peer-reviewed manuscripts on systematic reviews and meta-analysis in top ranked statistical and medical journals such as JASA, Biometrics, Biostatistics, SIM, SMMR, BMJ, Clinical Trials, JNCI, AIDS, Epidemiology and AJE. Dr. Chu's research on innovative statistical methods improve meta-analysis has been supported by 8 grants from FDA, AHRQ, NIAID, NIDCR and NLM as the principal investigator. Dr. Chu serves as an Associate Editor for Journal of the American Statistical Association, the American Journal of Epidemiology, and Statistics and Its Interface.

Dr. Yong Chen is Associate Professor of Biostatistics at University of Pennsylvania. He directs a Computing, Inference and Learning Lab at University of Pennsylvania, which focuses on integrating fundamental principles and wisdoms of statistics into quantitative methods for tackling key challenges in modern biomedical data. Dr. Chen is an expert in synthesis of evidence from multiple data sources, including systematic review and meta-analysis, distributed algorithms, and data integration, with applications to comparative effectiveness studies, health policy, and precision medicine. He is also working on developing methods to deal with suboptimal data quality issues in health system data, dynamic risk prediction, pharmacovigilance, and personalized health management. He has over 100 publications in a wide spectrum of methodological and clinical areas. He has been principal investigator on a number of grants, including R01s from the National Library of Medicine and National Institute of Allergy and Infectious Diseases, and Improving Methods for Conducting Patient-Centered Outcomes Research grant from Patient-Centered Outcomes Research Institute. He is an elected fellow of the Society for Research Synthesis Methodology, and the International Statistical Institute.

**SC02: Including historical data in clinical trial design and analysis****Length:** Half-day**Instructor:** Dr. Frank Fleischer (Boehringer-Ingelheim Pharma GmbH & Co. KG); Dr. Martin Oliver Sailer (Boehringer-Ingelheim Pharma GmbH & Co. KG)

**Outline/Description:** With the growing number of targeted drug development programs, there is an ever increasing interest to make these programs more cost effective. Borrowing of information from historical data allows to reduce the number of patients recruited to new trials and helps to bring new therapies to patients faster. Participants will learn requirements for the use of historical data in clinical trial design and analysis. It will be shown how Bayesian hierarchical models can be used to borrow information from historical data and perform Bayesian evidence synthesis with meta-analytic predictive priors. Advantages of dynamic weighting will be motivated. Since the population in the historical data and the new study may differ, propensity score methods and methods for covariate adjustment need to be considered. Case studies will be presented for examples from dose finding in oncology, basket trials and go/no-go decision making after phase II. Considerations for confirmatory settings will be addressed. Participants will be able to implement methods with computer exercises.

**About the Instructor:** Dr. Frank Fleischer Being a trained mathematician and statistician Frank has worked for more than 10 years in the pharmaceutical industry. He is heading a global team of statisticians at Boehringer Ingelheim focusing on statistical methodology and the implementation of innovative statistical designs into practice. In that role, Frank

and his team are considered with methodological questions regarding adaptive designs, statistical decision making, dose finding and Bayesian borrowing designs as well as with piloting these methods in clinical trials. Through this function several projects across different therapeutic areas and phases are supported. Formerly he has been a lead project statistician for different projects in oncology, immunology and the biosimilars.

Dr. Martin Oliver Sailer, With nine years of experience in the pharmaceutical industry, he has been Statistical lead for multiple pivotal Oncology and Biosimilar development programs. His consulting work focuses on introducing Bayesian methods in all phases of clinical development. His research interests include Design of Experiments, Bayesian Statistics, Basket designs, Statistical Go/No-Go decision making, and Estimands. He studied Statistics at TU Dortmund University in Germany and Iowa State University, Ames, IA.

### SC03: A Short Course on Absolute Risk Prediction

**Length:** Full-day

**Instructors:** Dr. Mitchell H Gail (National Cancer Institute, NIH trained at Harvard Medical School and in statistics at George Washington University, Department of Biostatistics, Johns Hopkins University); Dr. Ruth Pfeiffer (National Cancer Institute, graduate of the University of Maryland, College Park and the Technical University of Vienna, Austria Duration)

**Outline/Description:** Absolute (or “crude”) risk is the probability that an individual who is free of a given disease at an initial age,  $a$ , will develop that disease in the subsequent interval  $(a, t]$ . Absolute risk is reduced by mortality from competing risks. Models of absolute risk that depend on covariates have been used to design intervention studies, to counsel patients regarding their risks of disease and to inform clinical decisions. This course will define absolute risk and discuss methodological issues relevant to the development and evaluation of risk prediction models. Various study designs and data for model building will be presented, including cohort, nested case-control, and case-control data combined with registry data. Issues relating to the evaluation of risk prediction models and the strengths and limitations of risk prediction models for various applications will be discussed. Standard criteria for model assessment will be presented, as well as loss function-based criteria applied to the use of risk models to screen a population and the use of risk models to decide whether to take a preventive intervention that has both beneficial and adverse effects. Methods for validating models in independent data when some predictors are missing are presented. Finally, updating risk models when information on new predictors becomes available will be discussed.

**About the Instructors:** Dr. Mitchell H. Gail is a Senior Investigator at the Biostatistics Branch of the Division of Cancer Epidemiology and Genetics, National Cancer Institute

(NCI). Dr. Gail’s current research interests include statistical methods for the design and analysis of epidemiologic studies, and the development and application of models to predict the absolute risk of disease. Dr. Pfeiffer and Dr. Gail recently wrote a book entitled “Absolute Risk: Methods and Applications in Clinical Management and Public Health”. Dr. Gail served as President of the American Statistical Association and is a member of the National Academy of Medicine.

Dr. Ruth Pfeiffer is a tenured senior investigator at the Biostatistics Branch of the Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute (NCI). She received an M.S. degree in applied mathematics from the Technical University of Vienna, Austria, an M.A. degree in applied statistics and a Ph.D. in mathematical statistics both from the University of Maryland, College Park. At NCI she is an active collaborator on many research projects and mentors several fellows and junior investigators. Her research focuses on statistical methods for absolute risk prediction, problems arising in molecular and genetic epidemiologic studies and the analysis of data from electronic medical records. She is the recipient of a Fulbright Fellowship, an elected Member of the International Statistical Institute, and an elected Fellow of the American Statistical Association.

### SC04: Utilizing Real-World-Data and Real-World-Evidence in Drug Development and Evaluation

**Length:** Full-day

**Instructors:** Dr. Binbing Yu (AstraZeneca Oncology Biometrics); Dr. Bo Lu (Ohio State University (OSU), Division of Biostatistics College of Public Health); Dr. Qing Li (Takeda Pharmaceutical Company)

**Outline/Description:** In recent years, the rapid increase in the volume, variety, and accessibility of digitized RWD and RWE has presented unprecedented opportunities for the use of RWD and RWE throughout the drug product lifecycle. In clinical development, RWD and RWE have the potential to improve the planning and execution of clinical trials, and create a virtual control arm for a single arm for accelerated approval and label expansion. From the product lifecycle perspective, effective insights gleaned from RWE bring about informative relative benefits of drugs, comparative effectiveness, price optimization, and new indications. The goal of the short course is to serve as resources for practitioners who wish to apply these modern statistics and analytics in drug research and development. This short course will cover the essential statistical methodology for causal inference and recent practical case studies that adopted RWD and RWE in the clinical development and evaluation. In the morning session, we will introduce the current trend, challenges and opportunities of RWD and RWE in drug development and evaluation. We will also provide a comprehensive review of the relevant statistical methods for treatment effect estimation using non-randomized data, including propensity score match-

ing/stratification/weighting and sensitivity analysis. In the afternoon session, we will illustrate how to apply advanced statistical tools to practical case studies, including RWD and RWE in the clinical development and post-marketing drug development.

**About the Instructors:** Dr. Binbing Yu is an Associate Director in the Oncology Statistical Innovation group in AstraZeneca. He serves as the statistical expert across the whole spectrum of drug R&D process, including early-clinical and clinical research, design, operation and manufacturing, clinical pharmacology, oncology medical affairs and post-marketing surveillance. He obtained his PhD in Statistics from the George Washington University. His primary research interests are clinical trial design and analysis, cancer epidemiology, cause inference in observation studies, PK/PD modeling and Bayesian analysis. He was previously the Biometry Section Chief in the National Institute on Aging. He has nearly 80 publications in scientific and statistical journals and published a book on statistical methods on immunogenicity. Dr. Bo Lu is a Professor of Biostatistics in the College of Public Health, the Ohio State University. He obtained his PhD in Statistics from the University of Pennsylvania. His primary research interest covers causal inference with observational data, matching/weighting adjustment for complex designs including multiple treatment arms, time-varying treatment initiation, with complex survey weights, etc. Bayesian nonparametric modeling for heterogeneous causal effects, and statistical methods for survey sampling. He has been PIs for both federal and local-funded research grants on causal inference methodology. He has served as the lead statistician for the Ohio Medicaid Assessment Survey series since 2008. He also has extensive collaborations with Pharmaceutical industry on utilizing causal inference methods to leverage RWD in drug discovery.

Dr. Qing Li is a senior statistician in the statistical methodology group under the statistics and quantitative science (SQS) department at Takeda Pharmaceutical Company. His responsibilities include statistical methodology development and consultation for real-world-evidence (RWE) and advanced adaptive design from proof-of-concept to late phase studies across multiple therapeutic areas including oncology, gastroenterology (GI), rare disease, and vaccine. His research interests include propensity score (PS) methods, RWE, adaptive designs (sample size re-estimation, subgroup enrichment design, seamless design), Immuno-Oncology (IO) design and surrogate endpoints. He obtained his MS and PhD degree in biostatistics from the University of Iowa.

## SC05: Empower Statistician with Spark, Machine Learning and Deep Learning

**Length:** Full-day

**Instructors:** Dr. Hui Lin (Netflix); Dr. Ming Li (Amazon)

**Outline/Description:** Data can be a valuable asset, especially when there's a lot of it. Exploratory data analysis, busi-

ness intelligence, and machine learning can benefit tremendously if such big data can be wrangled and modelled at scale. Apache Spark is an open-source distributed engine for querying, processing and modeling big data. In this one-day workshop, you will learn how to leverage Spark and R/Python to process and model big data with common machine learning algorithm. By the end of this workshop, you will have a solid understanding of how to process big data using Spark and how to build common machine learning models in the cloud environment. You will also learn the motivation and use cases of deep learning through hands-on exercises. This workshop is designed for audience with statistics education background. This course bridges the gap between traditional statisticians and data scientists. No software download or installation is needed, everything is done through laptop's internet browser (Chrome or Firefox) with Databricks free cloud environment.

**About the Instructors:** Hui Lin is the head of data science at Netflix where she is leading and building the data science department. Before Netflix, she was a Data Scientist at DuPont. She provided data science leadership for a broad range of predictive analytics and market research analysis from 2013 to 2018. She is the co-founder of Central Iowa R User Group, blogger of <https://scientistcafe.com/>, and 2018 Program Chair of ASA Statistics in Marketing Section. She enjoys making analytics accessible to a broad audience and teaches tutorials and workshops for practitioners on data science (<https://course2019.scientistcafe.com/>). She holds MS and Ph.D. in statistics from Iowa State University.

Dr. Ming Li is currently a Research Scientist at Amazon. He organized and presented 2018 JSM Introductory Overview Lecture: Leading Data Science: Talent, Strategy, and Impact. He was the Chair of Quality & Productivity Section of ASA. He was a Data Scientist at Walmart and a Statistical Leader at General Electric Global Research Center before joining Amazon. He obtained his Ph.D. in Statistics from Iowa State University in 2010. With deep statistics background and a few years' experience in data science and machine learning, he has trained and mentored numerous junior data scientist with different backgrounds such as statistician, programmer, software developer, database administrator and business analyst. He is also an Instructor of Amazon's internal Machine Learning University and was one of the key founding members of Walmart's Analytics Rotational Program.

## SC06: Estimands and Statistical Methods for Missing data in Clinical Trials

**Length:** Half-day

**Instructor:** Dr. Frank Liu (Merck & Co.); Dr. Mandy Jin (AbbVie Inc.)

**Outline/Description:** In longitudinal clinical trials, data may be missing due to intercurrent events such as missing visits or early discontinuation. The strategies discussed in ICH E9 (R1) addendum for handling intercurrent events re-

quires clearly defined estimands and associated assumptions about missing data. To evaluate the underline treatment effects of an investigational new drug or biologics, it is desirable to consider estimands that can define an attributable causal inference for outcomes. Properly analyzing missing data with appropriate methods is critical to assess the attributable estimands. Commonly used approaches for missing data assume data are missing at random (MAR) and analyze data using likelihood-based methods or multiple imputations (MI). Because the MAR assumption is often difficult to justify, both regulatory agencies and industry sponsors have been seeking alternative approaches to handle missing data under missing not at random (MNAR) assumption, which estimates attributable estimands while excluding potential confounding. This half-day tutorial is intended to cover various methods that have been advocated in dealing with missing data and illustrates how to implement the analyses methods using examples. The tutorial begins with a review of estimands associated with missing data, followed by an overview of conventional methods for missing data handling such as maximum likelihood methods, multiple imputation, generalized estimation equation approaches, and Bayesian methods. The rest of the course is devoted to recently developed methods, including control-based imputation, tipping point analysis, and some methods developed by the instructors. Real clinical trial examples will be presented for illustration with implementation of the analysis using SAS software, including the MIXED, MI, MIANALYZE, GEE, and MCMC procedures.

**About the Instructor:** Dr. G. Frank Liu is a distinguished scientist at Merck & Co., Inc. and a Fellow of the American Statistical Association (ASA). For more than 24 years at Merck, Frank has gained extensive industry working experiences. His research interests include methods for longitudinal trials, missing data, safety analysis, and noninferiority trials; and has published more than 40 peer-reviewed statistical papers. He has been leading the development of many methodological guidance documents within Merck. He has taught short courses previously at Deming conferences, Biopharmaceutical Regulatory-Industry workshops, ASA conference on statistical practice, and conferences of the International Society of Biopharmaceutical Statistics.

Dr. Mandy Jin is currently a Director of Clinical Statistics at AbbVie Inc. She has gained 12 years of experience in clinical research across different therapeutic areas since she obtained her PhD in statistics from Columbia University in 2008. Her research interests include statistical methodologies for clinical trials, such as missing data, Bayesian analysis, adaptive designs, multiplicity adjustment, and machine learning. She has published more than 20 peer-reviewed statistical papers in these topics.

## SC07: Statistical Remedies for Flawed Conventions in Medical Research

**Length:** Half-day

**Instructor:** Dr. Peter F. Thall (The University of Texas MD Anderson Cancer Center)

**Outline/Description:** Many statistical methods commonly used for data analysis or clinical trial design by medical researchers are severely flawed. Unfortunately, some of these dysfunctional statistical conventions and paradigms are deeply embedded in the medical research community, and have become standard or even required practice. Ultimately, the consequence is that practicing physicians are misled to choose inferior or even harmful treatments for their patients. In this half day short course, I will identify and describe, by example, severe problems with a variety of statistical practices commonly used by medical statisticians and physician researchers. For each flawed practice, I will provide at least one practical alternative. Topics to be covered will include misinterpreting tests of hypotheses, misuse of p-values, evaluating strength of evidence, relationships between early treatment response and survival time, being misled by single-arm trials, futile futility rules, unsafe safety rules, Simpson's paradox, biomarkers and stratification, randomization and causality, bias correction, problems with outcome adaptive randomization, cherry picking, phase II-III designs, and dynamic treatment regimes.

**About the Instructor:** Dr. Peter F. Thall is the Anise J. Sorrell Professor in the Department of Biostatistics at M.D. Anderson Cancer Center. He is a Fellow of the American Statistical Association (ASA) and the Society for Clinical Trials, received the Don Owen Award in 2014, and is an ASA media expert.

Dr. Thall has published over 260 papers and book chapters in the statistical and medical literature, and co-authored the 2016 book *Bayesian Designs for Phase I-II Clinical Trials*. His latest book, *Statistical Remedies for Medical Researchers* will be published in early 2020. Dr. Thall's research areas include clinical trial design, precision medicine, Bayesian nonparametric statistics, incorporating expert opinion into Bayesian inference, and dynamic treatment regimes. He has presented over 200 invited talks and 30 short courses, and served as an associate editor for *Journal of the National Cancer Institute*, *Statistics in Medicine*, *Statistics in Biosciences*, *Clinical Trials*, and *Biometrics*.

## SC08: Statistics and Machine Learning Methods for EHR Data: From Data Extraction to Data Analytics/Predictions

**Length:** Full-day

**Instructor:** Dr. Hulin Wu (University of Texas Health Science Center at Houston, Center for Big Data in Health Sciences); Dr. Vahed Maroufy (School of Public Health, University of Texas Health Science Center-Houston (UTHealth));

Dr. Ashraf Yaseen (School of Public Health, University of Texas Health Science Center-Houston (UTHealth))

**Outline/Description:** This short course will provide an overview and present details of electronic health record (EHR) data extraction, cleaning, processing and analytics for scientific discoveries. The use of EHR data is becoming more prevalent for research purpose and deriving real-world evidence for decision or policy-making. However, analysis of this type of data has many unique complications due to how they are collected, processed, missing data issues, and types of questions that can be answered. This proposed short course covers many important topics related to using EHR data for research and scientific discoveries that include data extraction, cleaning, processing, making inference, and predictions based on many years of practical experience of instructors and their collaborators in the EHR Working Group at the University of Texas Health Science Center at Houston (UTHealth). Statistical and machine learning approaches will also be presented for EHR data extraction, cleaning and analysis. Additionally, since research projects for EHR Big Data are being conducted in large multidisciplinary research groups, the approaches for multiple-project management are necessary and will be also covered in this course.

**About the Instructor:** Dr. Wu joined the University of Texas Health Science Center at Houston (UTHealth) as Dr. D.R. Seth Family Professor and Associate Chair of Biostatistics and Professor of Biomedical Informatics in September 2015. He was appointed as the endowed Betty Wheless Trotter Professor and Chair for the newly named Department of Biostatistics & Data Science, UTHealth School of Public Health (SPH) in 2017. He is the Founding Director of the “Center for Big Data in Health Sciences” at UTHealth SPH with a goal to develop and use cutting-edge data science approaches to deal with Big Data from biomedical and health sciences. Dr. Wu was Dean’s Professor of Biostatistics and Computational Biology, Professor of Medicine, and Professor of Public Health Sciences at the University of Rochester Medical Center (URMC) from 2003-2015. He was the URMC Founding Director of the Center for Integrative Bioinformatics and Experimental Mathematics. Dr. Wu has extensive experience in directing NIH-funded research projects and contracts. As PI/Co-PI, he has been continuously funded by NIH since 1998 and he has received a total of \$30 million in NIH funding for independent research (R29 and 5 R01 grants), T32 training grant and NIH Cooperative Contract or center grants in the past 20 years. Dr. Wu has published 2 books and more than 130 peer-reviewed papers in statistics/biostatistics, biomathematics, bioinformatics and biomedical journals.

Dr. Maroufy is an Assistant Professor of Biostatistics at the department of Biostatistics and Data Science, School of Public Health-UTHealth. His research interests include data mining, statistical analysis and predictive modeling using big Electronic Health Records (EHR) and claim datasets. Currently his focus is on EHR data processing, cleaning, miss-

ing imputation and predictive analysis. Dr. Maroufy, has also experience and expertise in mathematical and methodological statistics such as mixture models, measurement error and sensitivity analysis using high-dimensional data.

Dr. Yaseen is currently an Assistant Professor of Data Science at the School of Public Health-UTHealth. His research interests include Machine Learning, Data Management & Analysis, Big Data, Bioinformatics, and High Performance Computing. In his current research work, Dr. Yaseen is exploring Big Data and Deep Learning technologies in Electronic Health Records data to address clinical and public health questions. He has extensive experience in computer programming, database design, implementation and management, web design and programming, and software engineering. He is actively contributing to several research projects at UTHealth for health-data analysis.

## SC09: Statistical Analysis of Microbiome Data with R

**Length:** Full-day

**Instructor:** Dr. Yinglin Xia (University of Illinois at Chicago); Dr. (Din) Ding-Geng Chen (University of North Carolina at Chapel Hill)

**Outline/Description:** Microbiome data are generated through either 16S rRNA gene sequencing or shotgun metagenomic sequencing. One unique feature of microbiome data is phylogenetic tree-structured. The bacterial taxa in a community are not randomly distributed; they usually not only depend on each other, but also exist the phylogenetic relationships among bacteria, which provides insights into the evolutionary relationships among bacterial taxa: a phylogenetic tree. Microbiome data have several features. The taxa abundance, amplicon sequence variants (ASVs) or operational taxonomic unit (OTU) counts, are naturally constrained, high dimensional, sparse with containing a large proportion of zero counts in the analysis data: feature table or OTU table. Typically, these data have complex covariance and correlation structures among different ASVs, OTUs, or taxa, and over-dispersed with large within-group heterogeneities. The unique data structure and all these data features pose the great challenges to analyze microbiome data using standard statistical methods and models. Recently we developed a statistical framework which consists of combining newly developed methods and models for microbiome data and borrowing methods and models from other fields such as ecology. This work was published in 2018 as a book titled “Statistical Analysis of Microbiome Data with R” by Springer (coauthored by Xia, Y., Sun, J. and Chen, D.G.) (<https://www.springer.com/us/book/9789811315336>). Since the book published in October, 2018, there are more than 40,000 downloads from Springer Bookmetrix, which is far more than the average downloads of statistical book from Springer. So far the readers from more than 30 countries have given us feedbacks and we were told that this book has been



used as textbook in Japan and several US universities. We were contacted frequently for requesting book material and slides for their teaching. The book review editor of the Biometrical Journal (Prof. and Dr. Annette Kopp-Schneider, the head of Division of Biostatistics, German Cancer Research Center, Germany) solicited a book review of this book, which published on 21 June 2019. This book was very positively reviewed (Biometrical Journal. 2019;1–2. [www.biometrical-journal.com](http://www.biometrical-journal.com) © 2019 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim DOI: 10.1002/bimj.201900176). Given the importance of microbiome study and currently only statistical book available, this book has been well received by peers of microbiome research. This course is designed to use this new book in this ICSA conference to meet the need of students and faculty to understand the microbiome data and perform the statistical analysis of microbiome data with R.

**About the Instructor:** Dr. Yinglin Xia is a Research Associate Professor at the Department of Medicine, the University of Illinois at Chicago, USA. He was a Research Assistant Professor in the Department of Biostatistics and Computational Biology at the University of Rochester, Rochester, NY. Dr. Xia has worked on a variety of research projects and clinical trials in microbiome, gastroenterology, oncology, immunology, psychiatry, sleep, neuroscience, HIV, mental health, public health, social and behavioral sciences, as well

as nursing caregiver. He has published more than 100 papers in peer-reviewed journals on Statistical Methodology, Clinical Trial, Medical Statistics, Biomedical Sciences, and Social and Behavioral sciences. He serves the editorial board for several scientific journals. He has successfully applied his statistical knowledge, modeling and programming skills to study designs and data analysis in biomedical research, clinical trials, and in microbiome research. He has written the first book, an invited review, and a book chapter on statistical analysis of microbiome data. He has designed four grants on microbiome studies funded by NIH, VA, and other funding agencies. His recent papers on microbiome data analysis are well received by peers.

Dr. Din Chen is a Fellow of ASA. He is now the Wallace H. Kuralt distinguished professor in biostatistics, University of North Carolina at Chapel Hill. He was a professor in biostatistics at the University of Rochester and the Karl E. Peace endowed eminent scholar chair in biostatistics at Georgia Southern University. Professor Chen is also a senior statistics consultant for biopharmaceuticals and government agencies with extensive expertise in clinical trials and bioinformatics. He has more than 150 referred professional publications and co-authored/co-edited 23 books on randomized clinical trials, statistical meta-analysis, public health statistical methods, causal inferences and statistical Monte-Carlo simulation and public health applications.

## Scientific Program (Dec. 13-16)

### Dec. 13 18:00 - 19:40

#### Session 1: Advancement of Machine Learning Methods via Tensors and High-Dimensional Tools

Organizer: Xuan Bi, University of Minnesota.

Chair: Xuan Bi, University of Minnesota.

- 18:00 High-order Joint Embedding for Multi-Level Link Prediction  
♦*Yubai Yuan and Annie Qu.* University of California, Irvine
- 18:25 Tensor denoising and completion based on ordinal observations  
*Chanwoo Lee and ♦Miaoyan Wang.* UW-Madison
- 18:50 Correlation Tensor Decomposition and Its Application in Spatial Imaging Data  
*Yujia Deng<sup>1</sup>, ♦Xiwei Tang<sup>2</sup> and Annie Qu<sup>3</sup>.* <sup>1</sup>University of Illinois Urbana-Champaign <sup>2</sup>University of Virginia <sup>3</sup>University of California Irvine
- 19:15 Unconventional regression paradigm for microbiome compositional data with phylogenetic tree structure  
*Gen Li.* University of Michigan
- 19:40 Floor Discussion.

#### Session 2: Latest development in latent variable models and genetics

Organizer: Xinyuan Song, Chinese University of Hong Kong.

Chair: Yingying Wei, Chinese University of Hong Kong.

- 18:00 BUSseq: a Bayesian Hierarchical Model Providing One-stop Services for scRNA-seq Data  
*Fangda Song, Ga Ming Angus Chan and ♦Yingying Wei.* Chinese University of Hong Kong
- 18:25 Longitudinal Structural Topic Models for Estimating Latent Health Trajectories using Administrative Claims Data  
*Mengbing Li and ♦Zhenke Wu.* University of Michigan, Ann Arbor
- 18:50 Latent variable modeling in biomarker studies  
*Zheyu Wang.* Johns Hopkins University
- 19:15 Accounting for correlated horizontal pleiotropy in two-sample Mendelian randomization using correlated instrumental variants  
*Qing Cheng and ♦Jin Liu.* Duke-NUS Medical School
- 19:40 Floor Discussion.

#### Session 3: New methods for joint analysis of survival and longitudinal data

Organizer: Qingning Zhou, University of North Carolina at Charlotte, Liang Zhu, The University of Texas Health Science Center at Houston.

Chair: Qingning Zhou, University of North Carolina at Charlotte.

- 18:00 Joint Analysis of Interval Censored Survival Time and Longitudinal Data  
*Di Wu and ♦Chenxi Li.* Michigan State University
- 18:25 Semiparametric latent-class models for multivariate longitudinal and survival data  
♦*Kin Yau Wong<sup>1</sup>, Donglin Zeng<sup>2</sup> and Dan-Yu Lin<sup>2</sup>.* <sup>1</sup>Hong Kong Polytechnic University <sup>2</sup>University of North Carolina at Chapel Hill
- 18:50 Regression analysis with proportional intensity model for general mixed recurrent event data with terminal events  
♦*Liang Zhu<sup>1</sup>, Yimei Li<sup>2</sup> and Gregory T. Armstrong<sup>3</sup>.* <sup>1</sup>The University of Texas Health Science Center at Houston <sup>2</sup>St. Jude Children's Research Hospital <sup>3</sup>St. Jude Children's Research Hospital
- 19:15 Cost-effective analysis for active surveillance versus nephron-sparing surgery for Bosniak III renal cysts: An application of multistate model  
*Xu Zhang.* The University of Texas Health Science Center at Houston
- 19:40 Floor Discussion.

#### Session 4: The Data are BIG and We are PRECISE: New Statistical Methods for Precision Medicine.

Organizer: Lu Wang, University of Michigan, Peng Zhang, University of Michigan Medical School.

Chair: Lu Wang, University of Michigan.

- 18:00 Bayesian nonparametric survival regression for optimizing precision dosing of intravenous busulfan in allogeneic stem cell transplantation  
*Peter Thall.* The University of Texas MD Anderson Cancer Center
- 18:25 Inferring Longitudinal antiretroviral drugs effects on depressive symptomatology in homogenous people with HIV  
♦*Yanxun Xu<sup>1</sup>, Wei Jin<sup>1</sup>, Yang Ni<sup>2</sup> and Leah Rubin<sup>1</sup>.* <sup>1</sup>Johns Hopkins University <sup>2</sup>Texas A&M University
- 18:50 Precision medicine for patients with chronic liver diseases through medical imaging  
*Peng Zhang.* University of Michigan
- 19:15 Kernel-Involved-Dosage-Decision Learning method for estimating dynamic dosage regimes  
♦*Ming Tang<sup>1</sup>, Matthew Schipper<sup>2</sup>, Theodore Lawrence<sup>2</sup> and Lu Wang<sup>2</sup>.* <sup>1</sup>Boehringer Ingelheim (China) investment Co. Ltd <sup>2</sup>University of Michigan
- 19:40 Floor Discussion.

#### Session 5: Decipher cell heterogeneity in high-throughput data analysis

Organizer: Ziyi Li, Emory University.

Chair: Ziyi Li, Emory University.

- 18:00 Choice of Scale in the Estimation of Cell-type Proportions  
♦*Johann Gagnon-Bartsch<sup>1</sup> and Gregory Hunt<sup>2</sup>.* <sup>1</sup>University of Michigan <sup>2</sup>College of William and Mary

18:25 In silico cell type deconvolution by integrative single-cell RNA-seq and bulk RNA-seq analysis

*Mingyao Li.* University of Pennsylvania

18:50 Tumor cell total mRNA expression shapes the molecular and clinical phenotype of cancer

*Shaolong Cao<sup>1</sup>, Jennifer R. Wang<sup>1</sup>, Shuangxi Ji<sup>1</sup>, Peng Yang<sup>1</sup>, Matthew D. Montierth<sup>1</sup>, Shuai Guo<sup>1</sup>, John Paul Shen<sup>1</sup>, Xiao Zhao<sup>1</sup>, Jingxiao Chen<sup>1</sup>, Alfonso Urbanucci<sup>2</sup>, Jonas Demeulemeester<sup>3</sup>, Peter Van Loo<sup>3</sup> and ♦Wenyi Wang<sup>1</sup>.* <sup>1</sup>The University of Texas MD Anderson Cancer Center <sup>2</sup>Oslo University Hospital <sup>3</sup>The Francis Crick Institute

19:15 Cell type-specific Expression Quantitative Trait Loci

*Little Paul<sup>1</sup>, Dan-Yu Lin<sup>2</sup>, Yun Li<sup>2</sup> and ♦Wei Sun<sup>1</sup>.* <sup>1</sup>Fred Hutchinson Cancer Research Center <sup>2</sup>University of North Carolina at Chapel Hill

19:40 Floor Discussion.

### Session 6: Recent Development on Heterogeneity Analysis

Organizer: Yang Li, Renmin University of China.

Chair: Yifan Sun, Renmin University of China.

18:00 Integrated quantile rank test for gene-level associations in sequencing studies

♦*Tianying Wang, Iuliana Ionita-Laza and Ying Wei.* Columbia University

18:25 Regression Trees for Interval-Censored Data

*Ce Yang, ♦Liqun Diao and Richard Cook.* University of Waterloo

18:50 Simultaneous prediction intervals for high-dimensional vector autoregressive model

*Mengyu Xu.* University of Central Florida

19:15 Floor Discussion.

### Session 7: Current Development in Experimental Designs and Its Applications

Organizer: Min-Qian Liu, Nankai University, Jianfeng Yang, Nankai University.

Chair: Yaping Wang, School of Statistics, East China Normal University.

18:00 On design orthogonality, projection uniformity and maximin distance for computer experiments

♦*Yaping Wang<sup>1</sup>, Fasheng Sun<sup>2</sup> and Hongquan Xu<sup>3</sup>.* <sup>1</sup>East China Normal University <sup>2</sup>Northeast China Normal University <sup>3</sup>UCLA

18:25 A method of constructing maximin distance designs

♦*Wenlong Li<sup>1</sup>, Min-Qian Liu<sup>1</sup> and Boxin Tang<sup>2</sup>.* <sup>1</sup>Nankai University <sup>2</sup>Simon Fraser University

18:50 Sequential good lattice point sets for computer experiments

♦*Xueru Zhang<sup>1</sup>, Yong-Dao Zhou<sup>1</sup>, Dennis Lin<sup>2</sup> and Min-Qian Liu<sup>1</sup>.* <sup>1</sup>Nankai university <sup>2</sup>Purdue University

19:15 Floor Discussion.

### Dec. 14 9:00 - 10:00

#### Session 8: Keynote speech

Organizer: Tony Sun, University of Missouri.

Chair: Tony Sun, University of Missouri.

9:00 Learning from COVID-19 Data in Wuhan, USA and the World on Transmission, Health Outcomes and Interventions  
*Xihong Lin.* Harvard University

### Dec. 14 10:20 - 12:00

#### Session 9: Bayesian Methodology and Applications for Complex Biomedical Data

Organizer: Liang Zhu, The University of Texas Health Science Center at Houston.

Chair: Xiaoyan Lin, University of South Carolina.

10:20 High dimensional mediation model for neuroimaging data analysis

*Xiaoqing Wang<sup>1</sup>, Yimei Li<sup>1</sup>, Wilburn Reddick<sup>1</sup>, Heather Conklin<sup>1</sup>, Amar Gajjar<sup>1</sup>, Cheng Cheng<sup>1</sup> and ♦Zhaohua Lu.* <sup>1</sup>St. Jude Children's Research Hospital

10:45 Bayesian inferences for panel count data and interval-censored data with nonparametric modeling of the baseline functions

♦*Lu Wang<sup>1</sup>, Xiaoyan Lin<sup>2</sup> and Lianming Wang<sup>2</sup>.* <sup>1</sup>Western New England University <sup>2</sup>University of South Carolina

11:10 Bayesian Latent Factor on Image Regression with Nonignorable Missing Data

*Xiaoqing (Jade) Wang.* St. Jude Children's Research Hospital

11:35 Bayesian Semiparametric Regression Analysis of Multivariate Panel Count Data

*Chunling Wang<sup>1</sup> and ♦Xiaoyan Lin<sup>2</sup>.* <sup>1</sup>University of South Carolina <sup>2</sup>University of South Carolina

12:00 Floor Discussion.

#### Session 10: New Challenges in Lifetime Data Analyses

Organizer: Yanqing Sun, University of North Carolina at Charlotte, Yu Shen, University of Texas MD Anderson Cancer Center.

Chair: Yanqing Sun, The University of North Carolina at Charlotte.

10:20 Analysis of the Time-Varying Cox Model for Cause-Specific Hazard Functions With Missing Causes

♦*Fei Heng<sup>1</sup>, Yanqing Sun<sup>2</sup>, Seunggeun Hyun<sup>3</sup> and Peter Gilbert<sup>4</sup>.* <sup>1</sup>University of North Florida <sup>2</sup>University of North Carolina at Charlotte <sup>3</sup>University of South Carolina Upstate <sup>4</sup>Fred Hutchinson Cancer Research Center

10:45 Regression Analysis of Mixed Panel Count Data with Dependent Terminal Events

*Guanglei Yu<sup>1</sup>, Liang Zhu<sup>2</sup>, ♦Yang Li<sup>3</sup>, Jianguo Sun<sup>4</sup> and Leslie Robison<sup>5</sup>.* <sup>1</sup>Eli Lilly and Company <sup>2</sup>The University of Texas Health Science Center at Houston <sup>3</sup>University of North Carolina at Charlotte <sup>4</sup>University of Missouri-Columbia <sup>5</sup>St. Jude Children's Research Hospital

- 11:10 Semiparametric Estimation of the Cure Fraction in Population-based Cancer Survival Analysis  
*Ennan Gu<sup>1</sup>, ♦Jiajia Zhang<sup>1</sup>, Wenbin Lu<sup>2</sup>, Lianming Wang<sup>3</sup> and Federico Felizzi<sup>4</sup>*. <sup>1</sup>University of South Carolina <sup>2</sup>North Carolina State University <sup>3</sup>University of South Carolina, SC <sup>4</sup>F. Hoffmann-La Roche Ltd, Basel
- 11:35 Benefit-harm Tradeoff in Individualized Treatment with Censored Data  
*Shuai Chen*. University of California, Davis
- 12:00 Floor Discussion.

### Session 11: Novel Semiparametric and Machine Learning Tools in Complex Observational Studies

Organizer: Jiwei Zhao, State University of New York at Buffalo.  
 Chair: Jing Wu, The University of Rhode Island.

- 10:20 Case-cohort Studies with Multiple Interval-censored Disease Outcomes  
 ♦*Qingning Zhou<sup>1</sup>, Jianwen Cai<sup>2</sup> and Haibo Zhou<sup>2</sup>*.  
<sup>1</sup>University of North Carolina at Charlotte <sup>2</sup>University of North Carolina at Chapel Hill
- 10:45 svReg: Structural Varying-coefficient Regression Identifies Individualized Relationship between Brain Regions and Motor Impairment in Huntington Disease  
 ♦*Rakheon Kim<sup>1</sup>, Samuel Mueller<sup>2</sup> and Tanya Garcia<sup>1</sup>*.  
<sup>1</sup>Texas A&M University <sup>2</sup>University of Sydney
- 11:10 Efficient Semiparametric Inference for Two-Phase Studies with Outcome and Covariate Measurement Errors  
*Ran Tao*. Vanderbilt University Medical Center
- 11:35 Floor Discussion.

### Session 12: Statistical Inference and Modeling for High-Dimensional and Complex Data Structure

Organizer: Wenbo Wu, The University of Texas at San Antonio.  
 Chair: Xiaoli Kong, Loyola University Chicago.

- 10:20 The Conditional Adaptive Lasso and Its Sufficient Variable Selection  
*Chenlu Ke*. Virginia Commonwealth University
- 10:45 Specification tests for covariance structures in high-dimensional statistical models  
*Xiao Guo<sup>1</sup> and ♦Cheng Yong Tang<sup>2</sup>*. <sup>1</sup>University of Science and Technology of China <sup>2</sup>Temple University
- 11:10 Subspace Estimation with Automatic Dimension and Variable Selection in Sufficient Dimension Reduction  
*Jing Zeng, ♦Qing Mai and Xin Zhang*. Florida State University
- 11:35 Pseudo Estimation in Regression  
 ♦*Wenbo Wu<sup>1</sup> and Xiangrong Yin<sup>2</sup>*. <sup>1</sup>University of Texas at San Antonio <sup>2</sup>University of Kentucky
- 12:00 Floor Discussion.

### Session 13: Artificial Intelligence and Causal Inference

Organizer: Li Luo, University of New Mexico.  
 Chair: Li Luo, University of New Mexico.

- 10:20 Neural causal network learning  
 ♦*Momiao Xiong, Tao Xu and Yuanyuan Liu*. The University of Texas Health Science Center at Houston
- 10:45 Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection  
*Qiyang Ge<sup>1</sup>, ♦Xuelin Huang<sup>2</sup>, Shenyang Fang<sup>2</sup>, Shicheng Guo<sup>3</sup>, Wei Lin<sup>4</sup> and Momiao Xiong<sup>1</sup>*. <sup>1</sup>The University of Texas Health Science Center at Houston <sup>2</sup>The University of Texas MD Anderson Cancer Center <sup>3</sup>University of Wisconsin-Madison <sup>4</sup>Fudan University
- 11:10 Conditional generative adversarial networks and variational autoencoders for individualized biomarker selection and treatment effect estimation  
 ♦*Shenyang Fang<sup>2</sup>, Qiyang Ge<sup>1,2</sup>, Shicheng Guo<sup>3</sup>, Yuanyuan Liu<sup>1</sup>, Jeffrey E. Lee<sup>2</sup>, Wei Lin<sup>4</sup> and Momiao Xiong<sup>1</sup>*.  
<sup>1</sup>The University of Texas Health Science Center at Houston <sup>2</sup>The University of Texas MD Anderson Cancer Center <sup>3</sup>University of Wisconsin-Madison <sup>4</sup>Fudan University
- 11:35 Artificial Intelligence and Causal Inference Inspired Methods for Forecasting the Spread of Covid-19 in the United States  
 ♦*Zixin Hu<sup>1</sup>, Qiyang Ge<sup>1</sup>, Shudi Li<sup>2</sup>, Eric Boerwinkle<sup>2</sup>, Wei Li<sup>1</sup>, Li Jin<sup>1</sup> and Momiao Xiong<sup>2</sup>*. <sup>1</sup>Fudan University <sup>2</sup>The University of Texas Health Science Center at Houston
- 12:00 Floor Discussion.

### Session 14: New advances in modern statistical modeling and testing

Organizer: Suojin Wang, Texas A&M University.  
 Chair: Suojin Wang, Texas A&M University.

- 10:20 Spatiotemporal Autoregressive Partially Linear Varying Coefficient Models  
 ♦*Shan Yu, Lily Wang and Lei Gao*. Iowa State University
- 10:45 A new bootstrap assisted stationarity test in the time domain  
 ♦*Lei Jin<sup>1</sup> and Suojin Wang<sup>2</sup>*. <sup>1</sup>Texas A&M Corpus Christi <sup>2</sup>Texas A&M
- 11:10 Comparison of Difference Based Variance Estimators for Partially Linear Models  
 ♦*Guoyi Zhang and Yan Lu*. University of New Mexico
- 11:35 Floor Discussion.

### Session 15: Robust Methods in Missing Data and Causal Inference

Organizer: Linbo Wang, University of Toronto.  
 Chair: Linbo Wang, University of Toronto.

- 10:20 Pattern graphs: a graphical approach to nonmonotone missing data problems  
 ♦*Yen-Chi Chen and Mauricio Sadinle*. University of Washington
- 10:45 The Promises of Parallel Outcomes  
*Ying Zhou, Dehan Kong and ♦Linbo Wang*. University of Toronto
- 11:10 Propensity Score Calibration with Missing-at-Random Data  
*Peisong Han*. University of Michigan

11:35 CCmed: Cross-condition mediation analysis for identifying robust trans-associations mediated by cis-gene

♦ *Fan Yang*<sup>1</sup>, *Kevin Gleason*<sup>2</sup>, *Jiebiao Wang*<sup>3</sup>, *Jubao Duan*<sup>4</sup>, *Xin He*<sup>2</sup>, *Brandon Pierce*<sup>2</sup> and *Lin Chen*. <sup>1</sup>University of Colorado Anschutz Medical Campus <sup>2</sup>University of Chicago <sup>3</sup>University of Pittsburgh <sup>4</sup>NorthShore University Health System

12:00 Floor Discussion.

### Session 16: Functional Data Analysis: Theory and Application

Organizer: Ruzong Fan, Georgetown University Medical Center (GUMC).

Chair: Chi-yang Chiu, University of Tennessee, Health Science Center.

10:20 Optimal Function-on-Function Regression with Interaction between Functional Predictors

*Honghe Jin*<sup>1</sup>, ♦ *Xiaoxiao Sun*<sup>2</sup> and *Pang Du*<sup>3</sup>. <sup>1</sup>University of Georgia <sup>2</sup>University of Arizona <sup>3</sup>Virginia Tech

10:45 Stochastic Functional Linear Models and Malliavin Calculus

♦ *Jin Zhou*<sup>1</sup>, *Weimiao Wu*<sup>2</sup>, *Chi-Yang Chiu*<sup>3</sup> and *Bingsong Zhang*<sup>4</sup>. <sup>1</sup>University of Arizona <sup>2</sup>Yale University <sup>3</sup>University of Tennessee, Health Science Center <sup>4</sup>Georgetown University Medical Center

11:10 A reproducing kernel Hilbert space framework for functional classification

♦ *Peijun Sang*<sup>1</sup>, *Adam Kashlak*<sup>2</sup> and *Linglong Kong*<sup>2</sup>. <sup>1</sup>University of Waterloo <sup>2</sup>University of Alberta

11:35 Floor Discussion.

### Session 17: Recent advances in multivariate and high-dimensional statistics

Organizer: Yunxiao Chen, London School of Economics and Political Science.

Chair: Yunxiao Chen, London School of Economics and Political Science.

10:20 Compound Sequential Change Point Detection in Multiple Data Streams

*Yunxiao Chen*<sup>1</sup> and ♦ *Xiaou Li*<sup>2</sup>. <sup>1</sup>London School of Economics and Political Science <sup>2</sup>University of Minnesota

10:45 Subtask Analysis of Process Data Through a Predictive Model

♦ *Xuening Tang*<sup>1</sup>, *Jingchen Liu*<sup>2</sup> and *Zhiliang Ying*<sup>2</sup>. <sup>1</sup>University of Arizona <sup>2</sup>Columbia University

11:10 Spectral clustering via adaptive layer aggregation for multi-layer networks

*Sihan Huang*<sup>1</sup>, ♦ *Haolei Weng*<sup>2</sup> and *Yang Feng*<sup>3</sup>. <sup>1</sup>Columbia University <sup>2</sup>Michigan State University <sup>3</sup>School of Global Public Health, New York University

11:35 Detection of Two-Way Outliers in Multivariate Data and Application to Cheating Detection in Educational Tests

♦ *Yunxiao Chen*, *Yan Lu* and *Irina Moustaki*. London School of Economics

12:00 Floor Discussion.

### Session 18: Big Data Analysis: New Directions and Innovation

Organizer: Wenhui Sheng, Marquette University, Guannan Wang, College of William and Mary.

Chair: Wenhui Sheng, Marquette University.

10:20 Sharp Inference on Selected Subgroups in Observational Studies with High Dimensional Covariates

♦ *Jingshen Wang*<sup>1</sup> and *Xinzhou Guo*<sup>2</sup>. <sup>1</sup>UC Berkeley <sup>2</sup>Harvard University

10:45 Sharp Optimality for High Dimensional Covariance Testing  
*Yumou Qiu*. Iowa State University

11:10 Statistical Inference for Mean Functions of 3D Functional Objects

♦ *Yueying Wang*<sup>1</sup>, *Xinyi Li*<sup>2</sup>, *Guannan Wang*<sup>3</sup>, *Li Wang*<sup>1</sup>, *Brandon Klinedinst*<sup>1</sup> and *Auriel Willette*<sup>1</sup>. <sup>1</sup>Iowa State University <sup>2</sup>University of North Carolina at Chapel Hill <sup>3</sup>College of William & Mary

11:35 Transformation and Integration of Microenvironment Microarray Data

*Gregory Hunt*. William & Mary

12:00 Floor Discussion.

### Session 19: Recent Trends of Innovative Methodologies and Applications in Rare Disease Clinical Trials

Organizer: Jian Zhu, Servier Pharmaceuticals, Bingming Yi, Vertex Pharmaceuticals.

Chair: Yeting Du, Servier Pharmaceuticals.

10:20 A simulation study to evaluate slope model with mixed-model repeated measure for rare disease

♦ *Tianle Hu* and *Lixi Yu*<sup>1</sup>. <sup>1</sup>Sarepta Therapeutics

10:45 Adaptive Endpoints Selection with Application in Rare Disease

♦ *Heng Xu*<sup>1</sup>, *Yi Liu*<sup>1</sup>, *Robert A. Beckman*<sup>2</sup>. <sup>1</sup>nektar therapeutics <sup>2</sup>Georgetown University Medical Center

11:10 Snapshot Matching: A Method for Borrowing from Longitudinal Historical Control Data

♦ *Yiyue Lou* and *Glen Laird*. Vertex Pharmaceuticals

11:35 BOIN12: Bayesian Optimal Interval Phase I/II Trial Design for Utility-Based Dose Finding in Immunotherapy and Targeted Therapies

♦ *Ying Yuan*, *Yahong Zhou*<sup>1</sup>, *Dianel Li*<sup>2</sup>, *Fangrong Yan*<sup>3</sup> and *Ying Yuan*<sup>1</sup>. <sup>1</sup>University of Texas MD Anderson Cancer Center <sup>2</sup>Bristol-Myers Squibb <sup>3</sup>China Pharmaceutical University

12:00 Floor Discussion.

### Session 20: Data Analysis and Application for High-Throughput Biotechnologies

Organizer: Xiaohua Zhang, University of Macau.

Chair: Dandan Wang, University of Macau.

10:20 A powerful genome-wide association test for complex diseases

*Linchen He* and ♦ *Yongzhao Shao*. New York University School of Medicine

- 10:45 High-throughput Computational Biology: It's Not Just About the Numbers  
*Robert Nadon*. McGill University
- 11:10 Untangle Clonal Evolution to Guide Precision Neuro-Oncology Via Data-Intensive Science  
*Jiguang Wang*. Hong Kong University of Science and Technology
- 11:35 Issues of z-factor and an approach to avoid them for quality control in high-throughput screening studies  
♦*Xiaohua Zhang<sup>1</sup>, Dandan Wang<sup>1</sup>, Shixue Sun<sup>1</sup> and Heping Zhang<sup>2</sup>*. <sup>1</sup>University of Macau <sup>2</sup>Yale University
- 12:00 Floor Discussion.

- Organizer: Liang Zhu, The University of Texas Health Science Center.  
Chair: Zhigang Zhang, Memorial Sloan Kettering Cancer Center.
- 14:00 Cancer immunotherapy trial design with long-term survivors  
♦*Jianrong Wu and Xue Ding*. University of Kentucky
- 14:25 Smooth Density Estimation Based on Interval-censored Data with Auxiliary Information  
♦*Qiang Zhao and Martin Schmidt*. Texas State University
- 14:50 Group sequential design for historical control trials using error spending functions  
*Jianrong Wu<sup>1</sup> and Yimei Li<sup>2</sup>*. <sup>1</sup>University of Kentucky <sup>2</sup>St. Jude children's research hospital
- 15:15 On testing the sub-distribution functions under competing risks.  
*Zhigang Zhang*. Memorial Sloan-Kettering Cancer Center
- 15:40 Floor Discussion.

**Session 21: Statistical Methods for Sports Data Analytics**

Organizer: Yishu Xue, Travelers Insurance.

Chair: Guanyu Hu, University of Connecticut.

- 10:20 A Bayesian Marked Spatial Point Processes Model for Basketball Shot Chart  
♦*Jieying Jiao, Jun Yan and Guanyu Hu*. University of Connecticut
- 10:45 Grouped Spatial Point Process Model: an Application for Basketball Shot Chart  
♦*Hou-Cheng Yang<sup>1</sup>, Yishu Xue<sup>2</sup> and Guanyu Hu<sup>3</sup>*. <sup>1</sup>Florida State University <sup>2</sup>Travelers Insurance <sup>3</sup>University of Connecticut
- 11:10 Modeling Quarterback Decision Making in the National Football League  
♦*Matthew Reyers and Tim Swartz*. Simon Fraser University
- 11:35 A Bayesian decision-theoretic approach to uncertain ranks and orderings: Comparing players and lineups  
♦*Andres Barrientos<sup>1</sup>, Deborshee Shen<sup>2</sup>, Garritt Page<sup>3</sup> and David Dunson<sup>4</sup>*. <sup>1</sup>Florida State University <sup>2</sup>Duke University <sup>3</sup>Brigham Young University <sup>4</sup>Duke University
- 12:00 Floor Discussion.

**Dec. 14 12:20 - 13:50****Session 22: Panel discussion: Leadership in Statistics and Data Science**

Organizer: Kelly Zou, Viatrix.

Chair: Kelly Zou, Viatrix.

Panelist	Affiliation
Haoda Fu	Eli Lilly and Company
Mengling Liu	NYU Langone Health
Jie Tang	Lotus clinical research LLC
Yuanjia Wang	Columbia University
Tian Zheng	Columbia University
Richard Zink	Lexitas Pharma Services, Inc.
Lauren Lee	Pfizer

**Dec. 14 14:00 - 15:40****Session 23: Innovative statistical methods for complex survival data and the applications****Session 24: Methods and applications in large and complex data**

Organizer: Qingcong Yuan, Miami University.

Chair: Wenbo Wu, The University of Texas at San Antonio.

- 14:00 High-Dimensional Rank-Based Inference  
♦*Xiaoli Kong<sup>1</sup> and Solomon Harrar<sup>2</sup>*. <sup>1</sup>Loyola University Chicago <sup>2</sup>University of Kentucky
- 14:25 High dimensional change point detection using generalized distance metrics  
*Shubhadeep Chakraborty and Xianyang Zhang*. Texas A&M University
- 14:50 A Divide and Conquer Algorithm of Bayesian Density Estimation  
*Ya Su*. University of Kentucky
- 15:15 Nonparametric Methods for Complex Multivariate Data: Asymptotics and Small Sample Approximations  
♦*Yue Cui<sup>1</sup> and Solomon Harrar<sup>2</sup>*. <sup>1</sup>Missouri State University <sup>2</sup>University of Kentucky
- 15:40 Floor Discussion.

**Session 25: Better Evidence Syntheses in Data Science**

Organizer: Haitao Chu, University of Minnesota Twin Cities.

Chair: Haitao Chu, University of Minnesota Twin Cities.

- 14:00 Galaxy plot: a new visualization tool of bivariate meta-analysis studies  
*Yong Chen*. University of Pennsylvania
- 14:25 Data fusion using summary versus individual data: relative efficiency for random-effects models  
*Dungang Liu*. University of Cincinnati
- 14:50 Estimating the Reference Range from a Meta-analysis  
♦*Lianne Siegel<sup>1</sup>, M. Hassan Murad<sup>2</sup> and Haitao Chu<sup>1</sup>*. <sup>1</sup>University of Minnesota <sup>2</sup>Mayo Clinic
- 15:15 Predictive treatment ranking in Bayesian network meta-analysis  
*Lifeng Lin*. Florida State University
- 15:40 Floor Discussion.

**Session 26: Deciphering Multi-omics Data: Statistical Models and Computational Approaches for Biology and Health**

Organizer: Hua Tang, Stanford University.

Chair: Catherine Tcheandjieu, Stanford University.

14:00 Cluster Ensemble and Batch Effect Correction Methods for Single Cell RNA-sequencing Data  
*Yun Li*. University of North Carolina

14:25 Fine association testing for whole genome sequencing data with knockoffs  
*Zihuai He*. Stanford University

14:50 Causal Inference for Heritable Phenotypic Risk Factors Using Heterogeneous Genetic Instruments  
♦ *Jingshu Wang*<sup>1</sup>, *Qingyuan Zhao*<sup>2</sup>, *Jack Bowden*<sup>3</sup>, *Gibran Hemani*<sup>4</sup>, *George D. Smith*<sup>5</sup>, *Dylan S. Small*<sup>6</sup> and *Nancy R. Zhang*. <sup>1</sup>University of Chicago <sup>2</sup>University of Cambridge <sup>3</sup>University of Exeter <sup>4</sup>University of Bristol <sup>5</sup>University of Bristol <sup>6</sup>University of Pennsylvania

15:15 HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity  
*Shuang Jiang*<sup>1</sup>, *Guanghua Xiao*<sup>2</sup>, *Andrew Koh*<sup>2</sup>, *Yingfei Chen*<sup>3</sup>, *Bo Yao*<sup>2</sup>, *Qiwei Li*<sup>4</sup> and ♦ *Xiaowei Zhan*. <sup>1</sup>Southern Methodist University <sup>2</sup>University of Texas Southwestern Medical Center <sup>3</sup>University of Texas <sup>4</sup>The University of Texas at Dallas

15:40 Floor Discussion.

14:25 Titration of T-cell Engager (TiTE): A new method for Phase 1 dose-finding design for systematic intra-subject dose escalation with application to T-cell Engagers  
♦ *Chenjia Xu*<sup>1</sup>, *Bin Zhuo*<sup>2</sup> and *Erik Rasmussen*<sup>2</sup>. <sup>1</sup>Indiana University <sup>2</sup>Amgen Inc.

14:50 uTPI: A Utility-Based Toxicity Probability Interval Design for Dose Finding in Phase I/II Trials  
*Ruitao Lin*. The University of Texas MD Anderson Cancer Center

15:15 Floor Discussion.

**Session 29: Novel clinical trial designs in the era of precision medicine and immunotherapy**

Organizer: Ruitao Lin, The University of Texas MD Anderson Cancer Center.

Chair: Ruitao Lin, The University of Texas MD Anderson Cancer Center.

14:00 TITE-BOIN-ET: Time-to-event Bayesian optimal interval design to accelerate dose-finding based on both efficacy and toxicity outcomes  
*Kentaro Takeda*. Astellas

14:25 Novel Early Phase Clinical Trial Designs for Cancer Therapeutic Vaccines  
*Chenguang Wang*. Johns Hopkins University

14:50 Hierarchical Bayesian Clustering Design of Multiple Biomarker Subgroups (HCOMBS)  
♦ *Jun (Vivien) Yin*, *Daniel Kang*<sup>1</sup> and *Qian Shi*<sup>2</sup>. <sup>1</sup>University of Iowa <sup>2</sup>Mayo Clinic

15:15 Floor Discussion.

**Session 27: Advanced Adaptive Enrichment Designs in Confirmative Clinical Trials**

Organizer: Bo Xu, Boston Biomedical, Inc.

Chair: Alex Dmitrienko, Mediana Inc.

14:00 A Case Study of Adaptive Population Enrichment Design in a Phase 3 Oncology Trial  
*Bo Jin*. Boston Biomedical, Inc

14:25 Complex multiplicity problems in adaptive designs with population selection.  
♦ *George Kordzakhia* and *Alex Dmitrienko*. FDA

14:50 Practical considerations for adaptive enrichment design implementation  
♦ *Jianchang Lin*, *Sheela Kolluri*, *Veronica Bunn* and *Rachael Liu*. Takeda Pharmaceuticals

15:15 Discussant  
*Jared Christensen*. NA

15:40 Floor Discussion.

**Session 28: New Challenges and Opportunities in Early-Phase Oncology Trials**

Organizer: Xuejing Wang, Eli Lilly and Company.

Chair: Xuejing Wang, Eli Lilly and Company.

14:00 Practical Considerations in Implementing Modern Dose-Escalation Methods from an Industry's Perspective  
♦ *Yuanyuan Bian*, *Aimee Wang* and *Wei Zhang*. Eli Lilly and Company

**Session 30: Statistical inference and practical issues in psychiatry**

Organizer: Zhezhen Jin, Columbia University.

Chair: Zhezhen Jin, Columbia University.

14:00 Statistical Ethics in Psychiatry  
*Jane Kim*. Stanford University

14:25 Latent Class Mediator  
*Haiqun Lin*. Rutgers University

14:50 LONGITUDINAL CANONICAL CORRELATION ANALYSIS  
♦ *Seonjoo Lee*<sup>1</sup>, *Jongwoo Choi*<sup>2</sup> and *Zhiqian Fang*<sup>1</sup>. <sup>1</sup>Columbia University and New York State Psychiatric Institute <sup>2</sup>New York State Psychiatric Institute

15:15 Deep Neural Network for Interval-Censored Survival Outcome Using Genetic Data, with an Application to Predict AD Progression  
*Tao Sun*<sup>1</sup> and ♦ *Ying Ding*<sup>2</sup>. <sup>1</sup>Renmin University <sup>2</sup>University of Pittsburgh, USA

15:40 Floor Discussion.

**Session 31: Recent development in dynamic historical data borrowing: methodology and application in clinical trials**

Organizer: Chenghao Chu, Vertex Pharmaceuticals.

Chair: Chenghao Chu, Vertex Pharmaceuticals.

- 14:00 A Dynamic Frequentist Approach of Historical Data  
♦*Bingming Yi and Chenghao Chu*. Vertex Biopharmaceuticals
- 14:25 The promises and compromises of dynamic borrowing in clinical trials  
♦*Xiaodong Luo and Hui Quan*. Sanofi
- 14:50 Adaptive Conditional Borrowing of Historical Data in Rare Disease Development  
♦*Yingying Liu<sup>1</sup>, Peng Sun<sup>1</sup>, Charlie Cao<sup>1</sup>, Bo Lu<sup>2</sup>, Ming-Hui Chen<sup>3</sup>, John Zhong<sup>4</sup>, Richard Foster, Susie Sinks<sup>1</sup>, Fan Wu<sup>1</sup> and Giulia Gambino<sup>1</sup>*. <sup>1</sup>Biogen <sup>2</sup>Ohio State University <sup>3</sup>University of Connecticut <sup>4</sup>REGENXBIO
- 15:15 Floor Discussion.

### Session 32: Bayesian Analysis of Complex Survey Data

Organizer: Cici Bauer, The University of Texas Health Science Center at Houston.

Chair: Luis Leon Novelo, The University of Texas Health Science Center in Houston.

- 14:00 Statistical Integration and Inference via Multilevel Regression and Poststratification  
*Yajuan Si*. University of Michigan
- 14:25 Fully Bayesian Estimation under Dependent and Informative Cluster Sampling  
*Luis Leon Novelo*. The University of Texas Health Science Center at Houston
- 14:50 Bayesian bivariate models for identifying common spatial patterns in small area estimation using survey data with weights  
*Cici Bauer*. The University of Texas Health Science Center at Houston
- 15:15 Locally Adaptive Shrinkage in Generalized Linear Models  
♦*Andrew Womack<sup>1</sup> and Daniel Taylor-Rodriguez<sup>2</sup>*. <sup>1</sup>Indiana University <sup>2</sup>Portland State University
- 15:40 Floor Discussion.

### Session 33: Multiple phenotypes, Pleiotropy and Mendelian Randomization

Organizer: Xiaofeng Zhu, Case Western Reserve University.

Chair: Xiaofeng Zhu, Case Western Reserve University.

- 14:00 Identifying pleiotropic loci between type 2 diabetes and prostate cancer  
♦*Debashree Ray and Nilanjan Chatterjee*. Johns Hopkins University
- 14:25 Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects  
♦*Guanghao Qi and Nilanjan Chatterjee*. Johns Hopkins University
- 14:50 Pleiotropy and Mendelian Randomization analysis using GWAS summary statistics  
*Xiaofeng Zhu*. Case Western Reserve university
- 15:15 Floor Discussion.

### Session 34: New methods of clinical trial designs and analyses and sample size re-estimation

Organizer: Gaohong Dong, iStats Inc.

Chair: Victoria Chang, BeiGene, Ltd..

- 14:00 Some Discussions on Sample Size Re-estimation  
♦*Victoria Chang, Jianfei Zheng and Yan Ma*. BeiGene
- 14:25 A Robust Design Approach for Clinical Trials with Potential Nonproportional Hazards: A Straw Man Proposal  
*Satrajit Roychoudhury*. Pfizer Inc.
- 14:50 Impact of censoring and follow-up time and use iwht non-proportional hazards  
*Gaohong Dong<sup>1</sup>, Bo Huang<sup>2</sup>, Yu-Wei Chang<sup>3</sup>, Yodit Seifu<sup>4</sup>, James Song<sup>3</sup> and David Hoaglin<sup>5</sup>*. <sup>1</sup>iStats Inc <sup>2</sup>Pfizer <sup>3</sup>BeiGene <sup>4</sup>Merck <sup>5</sup>U of Massachusetts
- 15:15 Beyond Bonferroni Correction  $\{e2\}_{i80\}_{i93\}$  Consistency of Evidence in Clinical Studies  
♦*Qian Li, Qiqi Deng<sup>1</sup> and Naitee Ting<sup>1</sup>*. <sup>1</sup>Boehringer Ingelheim Pharmaceuticals
- 15:40 Floor Discussion.

### Session 35: Utilization of RWE in Drug Development: Case Studies

Organizer: Zhaoyang Teng, Servier Pharmaceuticals, Qing Li, Takeda Pharmaceuticals.

Chair: Jian Zhu, Servier Pharmaceuticals.

- 14:00 Using RWD in Design and Analysis of Clinical Trials - Case Studies  
*Yanwei Zhang*. Takeda Pharmaceutical Company Limited
- 14:25 Actual example of using real world data for drug development  
*Küchiro Toyozumi*. Janssen Pharmaceutical K.K
- 14:50 Adjust survival estimates in the presence of treatment switching for HTA  
*Yuqing Xu<sup>1</sup>, Meijing Wu<sup>2</sup>, Weili He<sup>2</sup>, Qiming Liao<sup>3</sup> and Yabing Mai<sup>2</sup>*. <sup>1</sup>University of Wisconsin – Madison <sup>2</sup>Abbvie Inc. <sup>3</sup>ViiV Healthcare
- 15:15 Real-World Evidence in Regulatory Science  
*Joan Xie*. Seagen
- 15:40 Floor Discussion.

### Session 36: Challenges and developments in analyzing complex data

Organizer: Yuzhen Zhou, University of Nebraska-Lincoln, Yunlong Feng, University at Albany.

Chair: Yuzhen Zhou, University of Nebraska-Lincoln.

- 14:00 Forecasting with clustering based spatially varying autoregressive model  
*Sayli Pokal, Yuzhen Zhou and Trenton Franz*. University of Nebraska Lincoln
- 14:25 Bias-corrected estimation of functional-coefficient autoregressive models with measurement errors  
*Pei Geng*. Illinois State University



- 14:50 Bayesian compositional regression with structured priors for microbiome feature selection  
♦Liangliang Zhang, Yushu Shi, Robert Jenq, Kim-Anh Do and Christine Peterson. The University of Texas MD Anderson Cancer Center
- 15:15 Floor Discussion.

**Session 37: Materials Informatics**

Organizer: Meng Li, Rice University.  
Chair: Meng Li, Rice University.

- 14:00 Adaptive exploration and optimization of crystal structures  
♦Arvind Krishna, Huan Tran, Roshan Joseph and Rampi Ramprasad. Georgia Institute of Technology
- 14:25 Accounting for Location Measurement Error in Imaging Data with Application to Atomic Resolution Images of Crystalline Materials  
Matthew Miller, Matthew Cabral, Elizabeth Dickey, James Lebeau and ♦Brian Reich. North Carolina State University
- 14:50 Sparse inverse covariance estimation with graph constraints for identifying structures of high entropy alloys  
Xinrui Liu<sup>1</sup>, Changning Niu<sup>2</sup> and ♦Meng Li<sup>3</sup>. <sup>1</sup>Shandong Normal University <sup>2</sup>QuesTek Innovations LLC <sup>3</sup>Rice University
- 15:15 Floor Discussion.

**Dec. 14 16:00 - 17:40****Session 38: Methodological Advances for Harmonizing Genomics Data to Enable Reproducible Biomedical Research**

Organizer: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center.  
Chair: Tao Sun, University of Pittsburgh.

- 16:00 ComBat-seq: batch effect adjustment for RNA-seq count data  
Yuqing Zhang<sup>1</sup>, Giovanni Parmigiani<sup>2</sup> and ♦Evan Johnson<sup>3</sup>. <sup>1</sup>Gilead Sciences <sup>2</sup>Dana Farber Cancer Institute <sup>3</sup>Boston University
- 16:25 Improving Predictor Generalizability Using Multiple Studies with Differing Feature Sets  
Yujie Wu<sup>1</sup>, Boyu Ren<sup>2</sup>, Giovanni Parmigiani<sup>3</sup> and ♦Prasad Patil<sup>4</sup>. <sup>1</sup>Harvard T.H. Chan School of Public Health <sup>2</sup>McLean Hospital/Harvard Medical School <sup>3</sup>Dana-Farber Cancer Institute/Harvard T.H. Chan School of Public Health <sup>4</sup>Boston University School of Public Health
- 16:50 Depth Normalization of Small RNA Sequencing: Using Data and Biology to Select a Best Method  
Yannick Duren<sup>1</sup>, Johannes Lederer<sup>1</sup> and ♦Li-Xuan Qin<sup>2</sup>. <sup>1</sup>Ruhr University Bochum <sup>2</sup>Memorial Sloan Kettering Cancer Center
- 17:15 A robust normalization method for zero-inflated microbiome sequencing data  
Jun Chen. Mayo Clinic
- 17:40 Floor Discussion.

**Session 39: Statistical Method Development Motivated by Biomedical Data Challenges**

Organizer: Yu Shen, The University of Texas MD Anderson Cancer Center.

Chair: Yu Shen, The University of Texas MD Anderson Cancer Center.

- 16:00 Too many covariates and too few cases? - a comparative study  
♦Qingxia Chen, Hui Nian, Yuwei Zhu, Keipp Talbot, Marie Griffin and Frank Harrell. Vanderbilt University Medical Center
- 16:25 Quantifying Diagnostic Accuracy Improvement of New Biomarkers for Competing Outcomes  
Zheng Wang<sup>1</sup>, ♦Yu Cheng<sup>1</sup>, Eric Seaberg<sup>2</sup> and James Becker<sup>3</sup>. <sup>1</sup>University of Pittsburgh <sup>2</sup>Johns Hopkins University <sup>3</sup>University of Pittsburgh
- 16:50 Analysis of Generalized Semiparametric Mixed Varying-Coefficients Models for Longitudinal Data  
♦Yanqing Sun<sup>1</sup>, Li Qi<sup>2</sup>, Fei Heng<sup>3</sup> and Peter Gilbert<sup>4</sup>. <sup>1</sup>University of North Carolina at Charlotte <sup>2</sup>Sanofi, Bridgewater, U.S.A. <sup>3</sup>University of North Florida <sup>4</sup>University of Washington and Fred Hutchinson Cancer Research Center
- 17:15 Explained Variance Decompositions for Mediation Effect Sizes with Multiple Exposures  
♦Shanshan Zhao<sup>1</sup>, Yue Jiang<sup>2</sup> and Jason Fine<sup>3</sup>. <sup>1</sup>NIEHS/NIH <sup>2</sup>Duke University <sup>3</sup>University of North Carolina - Chapel Hill
- 17:40 Floor Discussion.

**Session 40: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences**

Organizer: Lili Yu, Georgia Southern University, Xinyan Zhang, Kennesaw State University.

Chair: Xinyan Zhang, Kennesaw State University.

- 16:00 Measuring Diagnostic Accuracy and Selecting Optimal Cut-points for K-class Diseases Based on Concordance and Discordance with Application  
♦Jing Kersey, Hani Samawi, Jingjing Yin, Haresh Rochani and Xinyan Zhang. Georgia Southern University
- 16:25 Pathway-Structured Predictive Modeling for Multi-Level Drug Response in Multiple Myeloma  
♦Xinyan Zhang<sup>1</sup>, Wenzhuo Zhuang<sup>2</sup> and Nengjun Yi<sup>3</sup>. <sup>1</sup>Kennesaw State University <sup>2</sup>Soochow University <sup>3</sup>University of Alabama at Birmingham
- 16:50 Application of Empirical Likelihood methods on bivariate Mean Residual Life function  
♦Ali Jinnah and Yichuan Zhao. Georgia State University
- 17:15 Generalized mean residual life models for case-cohort and nested case-control studies  
♦Peng Jin, Anne Zeleniuch-Jacquotte and Mengling Liu. New York University School of Medicine
- 17:40 Floor Discussion.

**Session 41: Statistical methods for complex human genetic data**

Organizer: Yuehua Cui, Michigan State University.

Chair: Yuehua Cui, Michigan State University.

- 16:00 A Kernel-Based Neural Network for High-dimensional Genetic Data Analysis  
♦*Qing Lu, Xiaoxi Shen and Xiaoran Tong.* University of Florida
- 16:25 A Bayesian Graphical model to delineate essential enhancer regulations using genome editing data  
*Hao Wang.* Michigan State University
- 16:50 Statistical inference of 3D genome structures and its applications in human genetics  
*Jianrong Wang.* Michigan State University
- 17:15 Multivariate partial linear varying coefficients model for genetic association studies with multiple longitudinal traits  
♦*Honglang Wang<sup>1</sup>, Jingyi Zhang<sup>2</sup> and Yuehua Cui<sup>3</sup>.*  
<sup>1</sup>Indiana University-Purdue University Indianapolis <sup>2</sup>Wells Fargo, Charlotte <sup>3</sup>Michigan State University
- 17:40 Floor Discussion.

#### Session 42: Recent advances in statistical methods for missing data, measurement error and biased sampling

Organizer: Baojiang Chen, The University of Texas Health Science Center at Houston.

Chair: Baojiang Chen, The University of Texas Health Science Center at Houston.

- 16:00 A New Bayesian Joint Model for Longitudinal Count Data with Many Zeros, Intermittent Missingness, and Dropout with Applications to HIV Prevention Trials  
*Jing Wu<sup>1</sup>, ♦Ming-Hui Chen<sup>2</sup>, Elizabeth Schifano<sup>2</sup>, Joseph Ibrahim<sup>3</sup> and Jeffrey Fisher<sup>2</sup>.* <sup>1</sup>University of Rhode Island <sup>2</sup>University of Connecticut <sup>3</sup>University of North Carolina at Chapel Hill
- 16:25 Bayesian nonparametrics for missing data in EHRs  
*Michael Daniels.* University of Florida
- 16:50 Semiparametric Generalized Linear Models for Analysis of Longitudinal Data with Biased Observation-level Sampling  
♦*Paul Rathouz<sup>1</sup> and Jacob Maronge<sup>2</sup>.* <sup>1</sup>University of Texas at Austin <sup>2</sup>University of Wisconsin-Madison
- 17:15 Variable Selection for Proportional Hazards Models with High Dimensional Covariates subject to Measurement Error  
♦*Baojiang Chen<sup>1</sup>, Ao Yuan<sup>2</sup> and Grace Yi<sup>3</sup>.* <sup>1</sup>The University of Texas Health Science Center at Houston <sup>2</sup>Georgetown University <sup>3</sup>University of Western Ontario
- 17:40 Floor Discussion.

#### Session 43: Statistical Learning Advancement for Inference with Complex Biomedical Data

Organizer: Feifei Xiao, University of South Carolina.

Chair: Jian Wang, The University of Texas MD Anderson Cancer Center.

- 16:00 Peel Learning for Pathway-related Outcome Prediction  
♦*Rui Feng<sup>1</sup>, Yuantong Li<sup>2</sup>, Mengying Yan<sup>3</sup> and Edward Cantu<sup>1</sup>.* <sup>1</sup>University of Pennsylvania <sup>2</sup>Purdue University <sup>3</sup>George Washington University
- 16:25 Using Statistical Learning to Promote Evidence-based Precision Health Care  
*Lu Wang.* University of Michigan

- 16:50 Multi-omic integration to reveal functional consequences of DNA alterations in tumor  
♦*Xiaoyu Song<sup>1</sup>, Jiayi Ji<sup>1</sup>, Lin Chen<sup>2</sup> and Pei Wang<sup>1</sup>.* <sup>1</sup>Icahn School of Medicine at Mount Sinai <sup>2</sup>University of Chicago
- 17:15 Microbial Network Recovery by Compositional Graphical Lasso  
*Chuan Tian, Duo Jiang, Thomas Sharpton and ♦Yuan Jiang.* Oregon State University
- 17:40 Floor Discussion.

#### Session 44: New Developments in High-Dimensional Data Analysis

Organizer: Xinyi Li, SAMSI.

Chair: Myungjin Kim, Iowa State University.

- 16:00 Sufficient dimension folding in regression via distance covariance for matrix-valued predictors  
♦*Wenhui Sheng<sup>1</sup> and Qingcong Yuan<sup>2</sup>.* <sup>1</sup>Marquette University <sup>2</sup>Miami University
- 16:25 Multivariate Dimension Reduction and the Dual Central Subspaces  
♦*Ross Iaci<sup>1</sup>, Xiangrong Yin<sup>2</sup> and Lixing Zhu<sup>3</sup>.* <sup>1</sup>The College of William and Mary <sup>2</sup>University of Kentucky <sup>3</sup>Hong Kong Baptist University
- 16:50 Generalized Spatially Varying Coefficient Models  
♦*Myungjin Kim and Li Wang.* Iowa State University
- 17:15 Spatial Autoregressive Partially Linear Varying Coefficient Models  
*Jingru Mu.* Kansas State University
- 17:40 Floor Discussion.

#### Session 45: Empirical Likelihood Methods and Bayesian Variable Selection

Organizer: Xinlei Wang, Southern Methodist University, Yichen Cheng, Georgia State University .

Chair: Xinlei Wang, Southern Methodist University.

- 16:00 Bayesian empirical likelihood based methods  
♦*Yichen Cheng and Yichuan Zhao.* Georgia State University
- 16:25 The Bayesian Elastic Net based on Empirical Likelihood  
*Adel Bedoui<sup>1</sup> and ♦Chul Moon<sup>2</sup>.* <sup>1</sup>Boehringer Ingelheim <sup>2</sup>Southern Methodist University
- 16:50 Reduce the computation in jackknife empirical likelihood for comparing two correlated Gini indices  
*Kangni Alemjrodo and ♦Yichuan Zhao.* Georgia State University
- 17:15 Full likelihood inference for abundance from capture-recapture data  
*Pengfei Li.* University of Waterloo
- 17:40 Floor Discussion.

#### Session 46: Statistical Process Control and Detection of Change-Point

Organizer: Dong Han, Shanghai Jiao Tong University.

Chair: Pang Du, Virginia Tech.

- 16:00 Variance Change Point Detection Under a Smoothly-Changing Mean Trend with Application to Liver Procurement  
♦Zhenguo Gao<sup>1</sup>, Zuofeng Shang<sup>2</sup>, Pang Du<sup>3</sup> and John Robertson<sup>3</sup>. <sup>1</sup>School of Mathematical Sciences, Shanghai Jiao Tong University <sup>2</sup>IUPUI <sup>3</sup>Virginia Tech
- 16:25 Optimal rate of convergence of multivariate nonparametric change point detection  
Xin Xing<sup>1</sup>, Zuofeng Shang<sup>2</sup>, ♦Pang Du<sup>3</sup>, Hongyu Miao<sup>4</sup> and Jun Liu<sup>1</sup>. <sup>1</sup>Harvard University <sup>2</sup>New Jersey Institute of Technology <sup>3</sup>Virginia Tech <sup>4</sup>The University of Texas Health Science Center at Houston
- 16:50 A Method of Optimizing the Control Charts for Finite Sequence of Observations  
♦Dong Han<sup>1</sup>, Fugee Tsung<sup>2</sup> and Jinguo Xian<sup>1</sup>. <sup>1</sup>Shanghai Jiao Tong University <sup>2</sup>Hong Kong University of science and technology
- 17:15 Fault Classification for High-dimensional Data Streams: A Directional Diagnostic Framework Based on Multiple Hypothesis Testing  
Dongdong Xiang. East China Normal University
- 17:40 Floor Discussion.

#### Session 47: Novel computational techniques for analyzing large scale biostatistical data

Organizer: Samiran Sinha, Texas A&M University, Tong Wang, Texas A&M University .  
Chair: Tong Wang, Texas A&M University.

- 16:00 Reduction of Bias Due to Misclassified Exposures using Instrumental Variables  
Christopher Manuel, ♦Samiran Sinha and Suojin Wang. Texas A&M University
- 16:25 Bayesian nonparametric bi-clustering of microbiome data  
Yang Ni. Texas A&M University
- 16:50 Consensus Monte Carlo for Random Subsets using Shared Anchors  
♦Peter Mueller<sup>1</sup>, Yang Ni<sup>2</sup> and Yuan Ji<sup>3</sup>. <sup>1</sup>UT Austin <sup>2</sup>TX A&M <sup>3</sup>U. Chicago
- 17:15 Connectivity Regression for heterogeneous networks  
Jeffrey Morris. University of Pennsylvania
- 17:40 Floor Discussion.

#### Session 48: Innovations in Statistical Machine Learning

Organizer: Boxiang Wang, University of Iowa.  
Chair: Yi Yang, McGill University.

- 16:00 Salient structure identification in complex networks by spectral periphery filtering  
♦Tianxi Li<sup>1</sup>, Elizaveta Levina<sup>2</sup> and Ji Zhu<sup>2</sup>. <sup>1</sup>University of Virginia <sup>2</sup>University of Michigan
- 16:25 Brain regions identified as being associated with verbal reasoning through the use of imaging regression via internal variation  
Long Feng<sup>1</sup>, ♦Xuan Bi<sup>2</sup> and Heping Zhang<sup>3</sup>. <sup>1</sup>City University of Hong Kong <sup>2</sup>University of Minnesota <sup>3</sup>Yale University

- 16:50 Penalized likelihood estimation under distance-to-set penalties via majorization-minimization  
Jason Xu. Duke University
- 17:15 Unsupervised Meets Supervised: Clustering of Regression Functions from Datasets  
♦Chenglong Ye<sup>1</sup> and Jie Ding<sup>2</sup>. <sup>1</sup>University of Kentucky <sup>2</sup>University of Minnesota
- 17:40 Floor Discussion.

#### Dec. 14 18:00 - 21:00

#### Session 49: Advances in Clinical Trial Statistics

Organizer: Poster organizers.  
Chair: Poster chairs.

- 18:00 Bayesian Optimal Phase II Design for Randomized Clinical Trials  
♦Yujie Zhao, Bo Yang, Jack J. Lee and Ying Yuan. The University of Texas MD Anderson Cancer Center
- 18:00 Elastic Meta-analytic-predictive Prior for Dynamically Borrowing Information from Historical Data with Application to Biosimilar Clinical Trials  
♦Wen Zhang<sup>1</sup>, Jean Pan<sup>2</sup> and Ying Yuan<sup>3</sup>. <sup>1</sup>The University of Texas Health Science Center at Houston <sup>2</sup>Amgen, Inc. <sup>3</sup>The University of Texas MD Anderson Cancer Center
- 18:00 Statistical Considerations in Clinical Trial Design with Event-free Survival as the Primary Efficacy Endpoint  
♦Yiming Zhang<sup>1</sup>, Tu Xu<sup>2</sup>, Meredith Goldwasser<sup>3</sup> and Vickie Zhang<sup>3</sup>. <sup>1</sup>University of Connecticut <sup>2</sup>Vertex Pharmaceuticals Inc. <sup>3</sup>Agios Pharmaceuticals Inc.
- 18:00 Trials of Targets  
♦Margret Erlendsdottir and Forrest Crawford. Yale School of Public Health
- 18:00 Analysis of crossover designs with nonignorable dropout  
♦Xi Wang and Chinchilli Vernon. Pennsylvania State University College of Medicine
- 18:00 Semiparametric isotonic regression analysis for risk assessment under nested case-control and case-cohort designs  
♦Wen Li<sup>1</sup>, Ruosha Li<sup>2</sup>, Ziding Feng<sup>3</sup> and Jing Ning<sup>4</sup>. <sup>1</sup>The University of Texas Health Science Center <sup>2</sup>The University of Texas Health Science Center at Houston <sup>3</sup>Fred Hutchinson Cancer Research Center <sup>4</sup>The University of Texas MD Anderson Cancer Center
- 18:00 TITE-BOIN12: A Bayesian Adaptive Design to Find the Optimal Biological Dose with Late-onset Toxicity and Efficacy  
♦Yanhong Zhou, Ruitao Lin, Jack Lee and Ying Yuan. The University of Texas MD Anderson Cancer Center
- 18:00 Floor Discussion.

#### Session 50: Bayesian Statistics

Organizer: Poster organizers.  
Chair: Poster chairs.

- 18:00 Bayesian Modeling of Spatial Transcriptomics Data via a Modified Potts Model  
♦*Xi Jiang*<sup>1</sup>, *Qiwei Li*<sup>2</sup> and *Guanghua Xiao*<sup>3</sup>. <sup>1</sup>Southern Methodist University <sup>2</sup>The University of Texas at Dallas <sup>3</sup>The University of Texas Southwestern Medical Center
- 18:00 A Bayesian Nonparametric Approach for Inferring Drug Combination Effects on Mental Health in People with HIV  
♦*Wei Jin*<sup>1</sup>, *Yang Ni*<sup>2</sup>, *Leah Rubin*<sup>3</sup>, *Amanda Spence*<sup>4</sup> and *Yanxun Xu*<sup>1</sup>. <sup>1</sup>Johns Hopkins University <sup>2</sup>Texas A&M University <sup>3</sup>Johns Hopkins University School of Medicine <sup>4</sup>Georgetown University
- 18:00 Informative sampling of Bayesian inference for continuous repeated measurement response  
*Helen Engle*. UTH
- 18:00 Evaluating Short-term Forecast among Different Epidemiological Models under a Bayesian Framework  
*Qiwei Li*<sup>1</sup>, ♦*Tejasv Bedi*<sup>1</sup>, *Guanghua Xiao*<sup>2</sup> and *Yang Xie*<sup>3</sup>. <sup>1</sup>University of Texas at Dallas <sup>2</sup>UT Southwestern Medical Center <sup>3</sup>UT Southwestern Medical Center.
- 18:00 Bayesian Landmark-Based Shape Analysis of Tumor Pathology Images  
♦*Cong Zhang*, *Kelli Palmer*, *Min Chen*, *Michael Zhang* and *Qiwei Li*. University of Texas at Dallas
- 18:00 BayesSMILES: Bayesian Segmentation Modeling for Longitudinal Epidemiological Studies  
♦*Shuang Jiang*<sup>1</sup>, *Quan Zhou*<sup>2</sup>, *Xiaowei Zhan*<sup>3</sup> and *Qiwei Li*<sup>4</sup>. <sup>1</sup>Southern Methodist University <sup>2</sup>Texas A&M University <sup>3</sup>University of Texas Southwestern Medical Center <sup>4</sup>The University of Texas at Dallas
- 18:00 Bayesian Functional Regression on Manifold With Application to Infant Cortical Thickness  
♦*Ye Emma Zohner*<sup>1</sup> and *Jeffrey Morris*<sup>2</sup>. <sup>1</sup>Rice University <sup>2</sup>University of Pennsylvania
- 18:00 Bayesian and Unsupervised Machine Learning Machines for Jazz Music Analysis  
*Qiuyi Wu*. University of Rochester
- 18:00 Double spike Dirichlet priors for structured weighting  
♦*Huiming Lin* and *Meng Li*. Rice University
- 18:00 Floor Discussion.

### Session 51: Recent developments in AI and its Applications

Organizer: Poster organizers.

Chair: Poster chairs.

- 18:00 Multi-scale affinities with missing data  
♦*Min Zhang*<sup>1</sup>, *Gal Mishne*<sup>2</sup> and *Eric Chi*<sup>1</sup>. <sup>1</sup>North Carolina State University <sup>2</sup>University of California San Diego
- 18:00 GPU accelerated statistical methods through a deep learning framework  
*Shikun Wang*. The University of Texas MD Anderson Cancer Center
- 18:00 Parameter Estimation and Inference of Spatial Autoregressive Model by Stochastic Gradient Descent  
♦*Gan Luan* and *Ji Meng Loh*. New Jersey Institute of Technology

- 18:00 Deep Learning for Quantile Regression: DeepQuantreg  
♦*Yichen Jia* and *Jong-Hyeon Jeong*. University of Pittsburgh
- 18:00 Algorithmic Regularized Fusion Minimization-Majorization Method for Clustering Histogram Data  
♦*Xu Han* and *Eric Chi*. North Carolina State University
- 18:00 Appearance-free Tripartite Matching for Multiple Object Tracking  
♦*Lijun Wang*<sup>1</sup>, *Yanting Zhu*<sup>2</sup>, *Jue Shi*<sup>2</sup> and *Xiaodan Fan*<sup>3</sup>. <sup>1</sup>Chinese University of Hong Kong <sup>2</sup>Hong Kong Baptist University <sup>3</sup>The Chinese University of Hong Kong
- 18:00 Recommender system of scholarly papers using public datasets  
♦*Jie Zhu*<sup>1</sup>, *Braja Patra*<sup>2</sup>, *Hulin Wu*<sup>1</sup> and *Ashraf Yaseen*<sup>1</sup>. <sup>1</sup>The University of Texas Health Science Center at Houston <sup>2</sup>The University of Texas Health Science Center at Houston; Weill Cornell Medicine
- 18:00 Extracting Clinically Meaningful Features for the Analysis of Tumor Pathology Images  
*Esteban Fernandezmorales*. The University of Texas at Dallas
- 18:00 Floor Discussion.

### Session 52: Statistics in Genetics

Organizer: Poster organizers.

Chair: Poster chairs.

- 18:00 Efficient odds ratio estimation using partial data audits in error-prone, observational HIV cohort data  
♦*Sarah C. Lotspeich*<sup>1</sup>, *Bryan E. Shepherd*<sup>2</sup>, *Gustavo G. C. Amorim*<sup>1</sup>, *Pamela A. Shaw*<sup>3</sup> and *Ran Tao*<sup>1</sup>. <sup>1</sup>Vanderbilt University <sup>2</sup>Vanderbilt University <sup>3</sup>University of Pennsylvania
- 18:00 Age-related alterations in fractal behaviors of respiratory signals  
♦*Teng Zhang*<sup>1</sup>, *Xinzheng Dong*<sup>2</sup>, *Chen Chang*<sup>1</sup> and *Xiaohua Douglas Zhang*<sup>1</sup>. <sup>1</sup>University of Macau <sup>2</sup>University of Macau, South China University of Technology
- 18:00 Something out of Nothing? The Influence of 0-0 Studies in Drug Safety Analysis  
♦*Zhaohu Fan*<sup>1</sup>, *Dungang Liu*<sup>1</sup>, *Yuejie Chen*<sup>2</sup> and *Nanhua Zhang*<sup>1</sup>. <sup>1</sup>University of Cincinnati <sup>2</sup>North Carolina State University
- 18:00 Epidemiology characteristics of influenza A and B in Macau  
*Hoiman Ng*<sup>1</sup>, *Teng Zhang*<sup>2</sup>, ♦*Guoliang Wang*<sup>2</sup>, *Simeng Kan*<sup>1</sup>, *Guoyi Ma*<sup>2</sup>, *Zhe Li*<sup>2</sup>, *Chang Chen*, *Dandan Wang*<sup>2</sup>, *Mengin Wong*<sup>1</sup> and *Chiohang Wong*<sup>1</sup>. <sup>1</sup>Kiang Wu Hospital <sup>2</sup>University of Macau
- 18:00 Prevalence of Allergen Sensitization in Patients with Allergic Diseases in mainland China: A Four-year Retrospective Study  
♦*Dandan Wang*<sup>1</sup>, *Wenting Luo*<sup>2</sup>, *Teng Zhang*<sup>1</sup>, *Peiyan Zheng*<sup>2</sup>, *Dongliang Leng*<sup>1</sup>, *Baoqing Sun*<sup>2</sup> and *Xiaohua Zhang*. <sup>1</sup>University of Macau <sup>2</sup>National Clinical Research Center of Respiratory Disease

- 18:00 Spatiotemporal profiling of COVID-19 epidemic in Hubei  
*Kuan Cheok Lei*. University of Macau
- 18:00 Covariate adjustment in continuous biomarker assessment  
♦ *Ziyi Li<sup>1</sup>, Zhenxing Guo<sup>1</sup>, Ying Cheng<sup>2</sup>, Peng Jin<sup>1</sup> and Hao Wu<sup>1</sup>*. <sup>1</sup>Emory University <sup>2</sup>Yunnan University
- 18:00 Floor Discussion.

### Session 53: Recent statistical advances in longitudinal and survival analysis

Organizer: Poster organizers.

Chair: Poster chairs.

- 18:00 Subgroup Analysis of Longitudinal Profiles for Compositional Count Data  
♦ *Chenyang Duan and Yuan Jiang*. Oregon State University
- 18:00 Comparative Analysis of Haplotype Assembly Algorithms  
♦ *Shuying Sun<sup>1</sup>, Flora Cheng<sup>2</sup>, Daphne Han<sup>3</sup>, Sarah Wei<sup>4</sup> and Alice Zhong<sup>5</sup>*. <sup>1</sup>Texas State University <sup>2</sup>Westwood High School <sup>3</sup>Kingwood High School <sup>4</sup>Massachusetts Institute of Technology <sup>5</sup>Clements High School
- 18:00 Transcriptome analysis reveals lncRNA-mediated complex regulatory network response to DNA damage in the liver tissue of *Rattus norvegicus*  
♦ *Chen Huang, Dong Liang Leng, Kuan Cheok Lei, Shi Xue Sun and Xiaohua Douglas Zhang*. University of Macau
- 18:00 Development of novel and robust method for analyzing single-cell RNA sequencing data  
♦ *Min Deng and Xiao Hua Douglas Zhang*. Faculty of Health Sciences
- 18:00 Expectile Neural Networks for Genetic Data Analysis of Complex Diseases  
♦ *Jinghang Lin<sup>1</sup>, Xiaoran Tong<sup>1</sup>, Chenxi Li<sup>1</sup> and Qing Lu<sup>2</sup>*. <sup>1</sup>Michigan State University <sup>2</sup>University of Florida
- 18:00 Floor Discussion.

### Session 54: Statistical innovations in medicine and public health

Organizer: Poster organizers.

Chair: Poster chairs.

- 18:00 On the Time-varying Predictive Performance of Longitudinal Biomarkers: Measure and Estimation  
*Jing Zhang<sup>1</sup>, ♦Jing Ning<sup>2</sup>, Xuelin Huang<sup>2</sup> and Ruosha Li<sup>1</sup>*. <sup>1</sup>The University of Texas Health Science Center at Houston <sup>2</sup>The University of Texas MD Anderson Cancer Center
- 18:00 New Families of Bivariate Copulas via Unit Weibull Distortion  
*Jungsywan Sepanski*. Central Michigan University
- 18:00 A family of partially linear single index models for analyzing complex environmental mixtures with continuous, categorical, survival, and longitudinal health outcomes  
♦ *Yuyan Wang, Yinxiang Wu, Myeonggyun Lee, Peng Jin, Leonardo Trasande and Mengling Liu*. NYU Langone Medical Center

- 18:00 Novel empirical likelihood inference for the mean difference with right-censored data  
♦ *Kangni Alemjdrodo and Yichuan Zhao*. Georgia State University
- 18:00 Semiparametric Marginal Regression Analysis for Clustered Competing Risks Data with Missing Cause of Failure  
♦ *Wenxian Zhou<sup>1</sup>, Giorgos Bakoyannis<sup>1</sup>, Ying Zhang<sup>2</sup> and Constantin Yiannoutsos<sup>1</sup>*. <sup>1</sup>Indiana University <sup>2</sup>University of Nebraska Medical Center
- 18:00 Floor Discussion.

### Session 55: Theory and Methodology for Big and Complex Data

Organizer: Poster organizers.

Chair: Poster chairs.

- 18:00 Robust Multiple Inference for Large-Scale Multivariate Regression  
♦ *Youngseok Song<sup>1</sup>, Wen Zhou<sup>1</sup> and Wen-Xin Zhou<sup>2</sup>*. <sup>1</sup>Colorado State University <sup>2</sup>University of California, San Diego
- 18:00 Revisiting Convexity-Preserving Signal Recovery with the Linearly Involved GMC Penalty  
♦ *Xiaoqian Liu and Eric Chi*. North Carolina State University
- 18:00 Diagnosing Learning Algorithms with Super-optimal Recursive Estimators  
♦ *Man Fung Leung and Kin Wai Chan*. Chinese University of Hong Kong
- 18:00 Semiparametric maximum likelihood estimation of panel count data with time-dependent covariates  
♦ *Dayu Sun<sup>1</sup> and Jianguo Sun<sup>2</sup>*. <sup>1</sup>Emory University <sup>2</sup>University of Missouri-Columbia
- 18:00 Regression Analysis of Multivariate Panel Count Data with Time-dependent Coefficient and Covariate Effects  
♦ *Yuanyuan Guo<sup>1</sup>, Dayu Sun<sup>2</sup> and Jianguo (Tony) Sun<sup>1</sup>*. <sup>1</sup>University of Missouri - Columbia <sup>2</sup>Emory University
- 18:00 Smoothed empirical likelihood for the difference of two quantiles with the paired sample  
♦ *Pangpang Liu and Yichuan Zhao*. Georgia State University
- 18:00 Latent Network Structure Learning from High Dimensional Multivariate Point Processes  
♦ *Biao Cai<sup>1</sup>, Jingfei Zhang<sup>2</sup> and Yongtao Guan<sup>1</sup>*. <sup>1</sup>Miami Herbert Business School <sup>2</sup>Miami Herbert School
- 18:00 Floor Discussion.

### Dec. 15 9:00 - 10:00

#### Session 56: Keynote speech

Organizer: Gang Li, Statistics and Decision Science, Janssen Research & Development.

Chair: Gang Li, Statistics and Decision Science, Janssen Research & Development.

- 9:00 Use of Real World Healthcare Data to Accelerate Vaccine Development in the Post COVID Era  
*Josh Chen*. Sanofi Pasteur

**Dec. 15 10:20 - 12:00****Session 57: Statistical Applications of Extreme Value Theory**

Organizer: Tiandong Wang, Texas A&M University.  
Chair: Tiandong Wang, Texas A&M University.

- 10:20 Semi-parametric estimation for multivariate extremes  
♦*John Nolan*<sup>1</sup>, *Anne-Laure Fougeres*<sup>2</sup> and *Cecile Mercadier*<sup>2</sup>. <sup>1</sup>American University <sup>2</sup>University of Lyon
- 10:45 All block maxima method for estimating the extreme value index  
*Jochem Oorschot* and ♦*Chen Zhou*. Erasmus University Rotterdam
- 11:10 Dynamic Bivariate Peak over Threshold Model for Joint Tail Risk Dynamics of Financial Markets  
*Zifeng Zhao*. University of Notre Dame
- 11:35 A Preferential Attachment Model with Poisson Growth  
♦*Tiandong Wang*<sup>1</sup> and *Sidney Resnick*<sup>2</sup>. <sup>1</sup>Texas A&M University <sup>2</sup>Cornell University
- 12:00 Floor Discussion.

**Session 58: Variable selection with complex lifetime data**

Organizer: Chenxi Li, Michigan State University.  
Chair: Chenxi Li, Michigan State University.

- 10:20 Simultaneously Variable Selection and Estimation for Interval-Censored Failure Time Data  
*Jianguo Sun*. University of Missouri
- 10:45 Learning survival from EMR/EHR data to estimate treatment effects using high dimensional claims codes  
*Ronghui Xu*. UC San Diego
- 11:10 Model Large-Scale Survival Data with Time-Varying Effects via a Minorization-Maximization Steepest Ascent Algorithm  
♦*Zhi (Kevin) He*, *Ji Zhu*, *Jian Kang* and *Yi Li*. University of Michigan
- 11:35 Variable selection for joint models with time-varying coefficients  
*Yujing Xie*<sup>1</sup>, ♦*Zangdong He*<sup>2</sup>, *Wanzhu Tu*<sup>3</sup> and *Zhangsheng Yu*<sup>1</sup>. <sup>1</sup>Shanghai Jiao Tong University <sup>2</sup>GlaxoSmithKline <sup>3</sup>Indiana University
- 12:00 Floor Discussion.

**Session 59: Recent advances in statistical methods for big biomedical data integration**

Organizer: Tianzhou Ma, University of Maryland, Mei-Ling Lee, University of Maryland.  
Chair: Tianzhou Ma, University of Maryland.

- 10:20 Outcome-guided Sparse K-means for Disease Subtype Discovery via Integrating Phenotypic Data with High-dimensional Transcriptomic Data  
*Lingsong Meng*<sup>1</sup>, *Dorina Avram*<sup>2</sup>, *George Tseng*<sup>3</sup> and ♦*Zhiguang Huo*<sup>1</sup>. <sup>1</sup>Department of Biostatistics, University of Florida <sup>2</sup>Department of Immunology, H. Lee Moffitt Cancer Center and Research Institute <sup>3</sup>Department of Biostatistics, University of Pittsburgh

- 10:45 Multiple testing correction for multivariate-multivariate association analysis: with application to an imaging-genetics study  
♦*Shuo Chen*<sup>1</sup> and *Qiong Wu*<sup>2</sup>. <sup>1</sup>University of Maryland <sup>2</sup>University of Maryland, College Park
- 11:10 Adaptive integration of testing results from multiple-trait genome-wide association studies  
♦*Chi Song* and *Qiaolan Deng*. Ohio State University
- 11:35 Combining p-values under arbitrary dependency structure in heavy-tailed distributions  
*Yusi Fang*<sup>1</sup>, *Chung Chang*<sup>2</sup>, *Yongseok Park*<sup>1</sup> and ♦*George Tseng*<sup>1</sup>. <sup>1</sup>University of Pittsburgh <sup>2</sup>National Sun Yat-Sen University
- 12:00 Floor Discussion.

**Session 60: Complex data analysis in business, economics, and industry**

Organizer: CY (Chor-yiu) Sin, National Tsing Hua University, Henghsiu Tsai, Academia Sinica, Taiwan.  
Chair: CY (Chor-yiu) Sin, National Tsing Hua University, Taiwan.

- 10:20 On Model Selection for ARFIMA and GARCH Processes  
*Ngai Hang Chan*<sup>1</sup>, ♦*Kun Chen*<sup>2</sup>, *Hsueh-Han Huang*<sup>3</sup> and *Ching-Kang Ing*<sup>3</sup>. <sup>1</sup>Chinese University of Hong Kong <sup>2</sup>Southwestern University of Finance and Economics <sup>3</sup>National Tsing Hua University
- 10:45 Testing for change points in heavy-tailed time series—A trimmed CUSUM approach  
*She Rui*<sup>1</sup> and ♦*Ling Shi*<sup>2</sup>. <sup>1</sup>Southwestern University of Finance and Economics <sup>2</sup>HKUST
- 11:10 A unified approach to bias approximations  
♦*Ruby Chiu-Hsing Weng*<sup>1</sup> and *Derek Stephen Coad*<sup>2</sup>. <sup>1</sup>National Chengchi University <sup>2</sup>Queen Mary, University of London
- 11:35 Model Averaging for High-dimensional Linear Regression Models with Dependent Observations  
♦*Ting-Hung Yu*<sup>1</sup>, *Ching-Kang Ing*<sup>2</sup> and *Henghsiu Tsai*<sup>3</sup>. <sup>1</sup>University of Iowa, U.S.A. <sup>2</sup>National Tsing Hua University <sup>3</sup>Academia Sinica
- 12:00 Floor Discussion.

**Session 61: Bayesian Analysis of Complex and High Dimensional Data**

Organizer: Xueying Tang, University of Arizona.  
Chair: Xueying Tang, University of Arizona.

- 10:20 Spatio-Temporal Additive Regression Model Selection for Urban Water Demand  
*Hunter Merrill*<sup>1</sup>, *Xueying Tang*<sup>2</sup> and ♦*Nikolay Bliznyuk*<sup>1</sup>. <sup>1</sup>University of Florida <sup>2</sup>University of Arizona
- 10:45 Joint Bayesian Analysis of Multiple Response-Types Using the Hierarchical Generalized Transformation Model  
*Jonathan Bradley*. Florida State University
- 11:10 Joint Bayesian Variable and DAG Selection Consistency for High-dimensional Regression Models with Network-structured Covariates  
♦*Xuan Cao*<sup>1</sup> and *Kyoungjae Lee*<sup>2</sup>. <sup>1</sup>University of Cincinnati <sup>2</sup>Inha University

- 11:35 Shrinkage on Simplex : Sparsity-Inducing Priors for Compositional Data  
*Jyotishka Datta*. University of Arkansas
- 12:00 Floor Discussion.

- 10:45 Joint model of temporal microbiome and risk of outcomes at matched time  
♦ *Qian Li*<sup>1</sup>, *Kendra Vehik*<sup>1</sup>, *Jeffery Krischer*<sup>1</sup> and *Yijuan Hu*<sup>2</sup>.  
<sup>1</sup>University of South Florida <sup>2</sup>Emory University

- 11:10 Integrative Analysis of Multi-Omic Data via Sparse Multiple Co-Inertia Analysis  
*Eun Jeong Min* and ♦ *Qi Long*. University of Pennsylvania

- 11:35 WEVar: a novel statistical learning framework for predicting noncoding regulatory variants  
*Li Chen*. Indiana University School of Medicine
- 12:00 Floor Discussion.

### Session 62: New statistical methods for machine learning on big data

Organizer: Jun Li, University of Notre Dame.  
Chair: Jun Li, University of Notre Dame.

- 10:20 Adversarial Machine Learning – Game Theoretic Approach and Adversarial Attack Against Deep Neural Networks  
*Bowei Xi*. Purdue University
- 10:45 A universal event detection framework for Neuropixels data  
♦ *Hao Chen*<sup>1</sup>, *Shizhe Chen*<sup>1</sup> and *Xinyi Deng*<sup>2</sup>. <sup>1</sup>University of California, Davis <sup>2</sup>Beijing University of Technology
- 11:10 Improved double robust approach for precision medicine  
*Lingsong Zhang*. Purdue University
- 11:35 A new clustering algorithm for assigning cells to known cell types according to marker genes  
*Hongyu Guo* and ♦ *Jun Li*. University of Notre Dame
- 12:00 Floor Discussion.

### Session 65: Statistical learning with complex data structure

Organizer: Tianxi Li, University of Virginia.  
Chair: Tianxi Li, University of Virginia.

- 10:20 Quantile Trend Filtering  
♦ *Eric Chi*<sup>1</sup>, *Halley Brantley*<sup>2</sup> and *Joseph Guinness*<sup>3</sup>.  
<sup>1</sup>North Carolina State University <sup>2</sup>United Health Group <sup>3</sup>Cornell University
- 10:45 Network Inference from Grouped Observations  
♦ *Yunpeng Zhao*<sup>1</sup>, *Peter Bickel*<sup>2</sup> and *Charles Weko*<sup>3</sup>.  
<sup>1</sup>Arizona State University <sup>2</sup>University of California, Berkeley <sup>3</sup>US Army
- 11:10 A Flexible Latent Space Model for Multilayer Networks  
♦ *Xuefei Zhang*, *Songkai Xue* and *Ji Zhu*. University of Michigan
- 11:35 Floor Discussion.

### Session 63: Innovative statistical methods for optimal treatment selection and clinical trial design with historical data

Organizer: Alan Wu, Celgene, Jing Gong, Celgene.  
Chair: Jing Gong, Celgene.

- 10:20 Machine Learning of Non-Randomized Control Studies for Causal Inference  
♦ *Xiaolong Luo*, *Mingyu Li*, *Jing Gong*, *Marie-Laure Casadebaig*, *Daniel Li* and *Mike Branson*. Bristol Meyers Squibb
- 10:45 Evaluation of different analytic strategies for estimating optimal treatment regimens for time-to- event outcomes in observational data  
♦ *Ilya Lipkovich*, *Zbigniew Kadziola*, *Bohdana Ratitch*, *Zhanglin Cui* and *Douglas Faries*. Eli Lilly and Company
- 11:10 Using Real-World Evidence at FDA/CBER  
*Jiang (Jessica) Hu*. NA
- 11:35 Discussant  
*Jingjing Ye*. BeiGene
- 12:00 Floor Discussion.

### Session 66: Bridging the gap between complex data and public health policies: methods and applications

Organizer: Zehang Li, Yale School of Public Health.  
Chair: Zehang Li, Yale School of Public Health.

- 10:20 VC-BART: Varying Coefficients  
♦ *Sameer Deshpande*<sup>1</sup>, *Ray Bai*<sup>2</sup>, *Cecilia Balocchi*<sup>2</sup> and *Jennifer Starling*<sup>3</sup>. <sup>1</sup>MIT <sup>2</sup>University of Pennsylvania <sup>3</sup>The University of Texas at Austin
- 10:45 Bounds for local average treatment effects in instrumental variable analyses of mobile interventions  
*Andrew Spieker*. Vanderbilt University Medical Center
- 11:10 Integrating Sample Relatedness Information into Latent Class Models: A Tree-Structured Shrinkage Approach  
*Zhenke Wu*. University of Michigan, Ann Arbor
- 11:35 Bayesian Latent Class Models for Verbal Autopsy Data from Multiple Domains.  
*Zehang Li*. University of California, Santa Cruz
- 12:00 Floor Discussion.

### Session 64: Statistical method advancement for analyzing omics data

Organizer: Hao (Harry) Feng, Case Western Reserve university, Xiaofeng Zhu, Case Western Reserve University.  
Chair: Hao (Harry) Feng, Case Western Reserve University.

- 10:20 Robust partial reference-free cell composition estimation from tissue expression  
♦ *Ziyi Li*<sup>1</sup>, *Zhenxing Guo*<sup>1</sup>, *Ying Cheng*<sup>2</sup>, *Peng Jin*<sup>1</sup> and *Hao Wu*<sup>1</sup>. <sup>1</sup>Emory University <sup>2</sup>Yunnan University

### Session 67: Advanced Bayesian methods in Biostatistics

Organizer: Michele Guindani, University of California, Irvine.  
Chair: Michele Guindani, University of California, Irvine.

- 10:20 A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes  
♦ *Matthew Koslovsky*<sup>1</sup> and *Marina Vannucci*<sup>2</sup>. <sup>1</sup>Rice University <sup>2</sup>Rice University

- 10:45 Graphical Models for Data Integration and Mediation Analysis  
*Min Jin Ha.* The University of Texas MD Anderson Cancer Center
- 11:10 Flexible and informative clustering of microbiome data  
*Christine Peterson.* The University of Texas MD Anderson Cancer Center
- 11:35 Dependent Mixtures: Modeling Cell Lineage  
♦*Giorgio Paulon, Carlos Paganizani and Peter Müller.* The University of Texas at Austin
- 12:00 Floor Discussion.

- 14:00 PA-CRM: A Continuous Reassessment Method for Pediatric Phase I Trials with Concurrent Adult Trials  
♦*Yimei Li<sup>1</sup> and Ying Yuan<sup>2</sup>.* <sup>1</sup>University of Pennsylvania <sup>2</sup>The University of Texas MD Anderson Cancer Center
- 14:25 A Review of the experience of pediatric written requests issued for oncology drug products  
*Jingjing Ye.* BeiGene
- 14:50 BOIN12: Bayesian Optimal Interval Phase I/II Trial Design for Utility-Based Dose Finding in Immunotherapy and Targeted Therapies  
*Ruitao Lin<sup>1</sup>, Yahong Zhou<sup>1</sup>, Dianel Li<sup>2</sup>, Fangrong Yan<sup>3</sup> and Ying Yuan<sup>1</sup>.* <sup>1</sup>The University of Texas MD Anderson Cancer Center <sup>2</sup>Bristol-Myers Squibb <sup>3</sup>China Pharmaceutical University
- 15:15 Floor Discussion.

### Session 68: Leveraging Real-World Data in Comparative Effectiveness Research

Organizer: Zonghui Hu, National Institute of Allergy and Infectious Diseases, Zhiwei Zhang, National Cancer Institute.  
Chair: Misrak Gezmu, National Institute of Allergy and Infectious Diseases.

- 10:20 Real world data, machine learning and causal inference  
*Jie Chen.* Overland Pharma
- 10:45 Estimation and variable selection for conditional causal effect: a dimension reduction approach  
*Zonghui Hu.* National Institutes of Health
- 11:10 Multilevel-Multiclass Graphical Model for Correlated Network  
*Inyoung Kim.* Virginia Tech
- 11:35 Adjusting for Population Differences Using Machine Learning Methods  
*Zhiwei Zhang.* National Institutes of Health
- 12:00 Floor Discussion.

### Dec. 15 12:20 - 13:50

#### Session 69: Panel discussion: Real World Evidence to Advance Healthcare

Organizer: Kelly Zou, Viatrix.  
Chair: Kelly Zou, Viatrix.

Panelist	Affiliation
Joseph Imperato	Viatrix
Joseph Cook	Viatrix
Aaron Galaznik	Medidata Solutions
Max Ma	Johnson and Johnson
Jun Su	Astellas US LLC
May Yamada-Lifton	SAS Institute

### Dec. 15 14:00 - 15:40

#### Session 70: Design and Statistical Issues for Pediatric Oncology Trials

Organizer: Haitao Pan, St. Jude Children's Research Hospital, Ying Yuan, University of Texas MD Anderson Cancer Center.  
Chair: Haitao Pan, St. Jude Children's Research Hospital.

#### Session 71: Enhancing RCT using Real World Evidence

Organizer: Meijing Wu, AbbVie Inc., Zailong Wang, AbbVie Inc..  
Chair: Jiewei Zeng, AbbVie Inc..

- 14:00 Creating A Synthetic Control Arm Using Propensity Score Analysis  
♦*Zailong Wang, Zhuqing Yu and Lanju Zhang.* ABBVIE
- 14:25 Use of Hybrid Control Arms with Adaptive Power Priors for Time-to-Event Data  
*Matthew Psioda.* UNC-CH
- 14:50 Analytic framework for non-randomized single-arm clinical trials with external RWD control  
♦*Hongwei Wang, Yixin Fang and Weili He.* AbbVie
- 15:15 Practical Considerations in Design an Execution of RWE Studies  
*Yijie Zhou.* Vertex Pharmaceuticals
- 15:40 Floor Discussion.

#### Session 72: Real World Evidence for Value-Added Patient-Centric Healthcare

Organizer: Kelly Zou, Viatrix, Jim Li, Viatrix .  
Chair: Kelly Zou, Viatrix.

- 14:00 National Health and Wellness Survey exploratory cluster analysis of males 40-70 years old focused on erectile dysfunction and associated risk factors across the USA, Italy, Brazil and China  
*Irwin Goldstein<sup>1</sup>, Amir Goren<sup>2</sup>, Ryan Liebert<sup>2</sup>, Wing Yu Tang<sup>3</sup> and ♦Tarek Hassan<sup>4</sup>.* <sup>1</sup>Alvarado Hospital <sup>2</sup>Kantar <sup>3</sup>Pfizer <sup>4</sup>Viatrix
- 14:25 Evidence and Access with Uncertainty: Establishing Value with Evolving Evidence  
♦*Joseph Cook<sup>1</sup> and Adam Heathfield<sup>2</sup>.* <sup>1</sup>Viatrix <sup>2</sup>Pfizer
- 14:50 Real World Evidence on the Impact of Sertraline Daily Treatment Regimen on Medication Adherence and Persistence in Patients with Major Depressive Disorder or Obsessive-Compulsive Disorder  
*Gang Wang<sup>1</sup>, Tianmei Si<sup>2</sup>, ♦Joseph Imperato<sup>3</sup>, Li Li Yang<sup>3</sup>, Kelly Zou<sup>3</sup>, Ying Jin<sup>3</sup>, Elizabeth Pappadopulos<sup>3</sup>, Lei Yan<sup>3</sup> and Wei Yu<sup>3</sup>.* <sup>1</sup>Beijing Anding Hospital Affiliated to Capital Medical University <sup>2</sup>Peking University <sup>3</sup>Viatrix



15:15 Harnessing real-world evidence to reduce the burden of non-communicable disease: health information technology and innovation to generate insights

*Kelly Zou*<sup>1</sup>, ♦ *Jim Li*<sup>1</sup>, *Lobna A.salem*<sup>1</sup>, *Joseph Imperato*<sup>1</sup>, *Jon Edwards*<sup>2</sup> and *Amrit Ray*<sup>3</sup>. <sup>1</sup>Viatris <sup>2</sup>Envision Pharma Group <sup>3</sup>Pfizer

15:40 Floor Discussion.

14:00 Statistical Challenges In CAR-T Cell Therapy Development  
*Daniel Li*. BMS

14:25 Statistical review of gene and cell therapies and related research topics

*Xue (Mary) Lin*. The Food and Drug Administration

14:50 Interpretation of the Treatment Effect in the Context of Complex Treatment Strategy and Methodological considerations for CAR-T clinical trials

*YiYun (Michael) Zhang*. Autolus Ltd

15:15 Statistics Considerations to Find Optimal Dose of CAR-T Cell Therapy

*Xiaoling Wu*. Legend Biotech

15:40 Floor Discussion.

### Session 73: High-dimensional statistical learning in big-data of human genetics

Organizer: Jianrong Wang, Science and Engineering, Michigan State University.

Chair: Jianrong Wang, Michigan State University.

14:00 Gene Network Analysis with Single Cell RNA Sequencing Data

♦ *Fei Zou and Meichen Dong*. UNC-CH

14:25 Multi-omics data integration with kernel fusion

*Haitao Yang*<sup>1</sup> and ♦ *Yuehua Cui*<sup>2</sup>. <sup>1</sup>Hebei Medical University <sup>2</sup>Michigan State University

14:50 MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations

♦ *Zifan Zhu, Jie Ren, Sonia Michail and Fengzhu Sun*. University of Southern California, Los Angeles

15:15 Joint modeling of bacterial and fungal network in Alcoholic hepatitis

*Xinlian Zhang*. University of California San Diego

15:40 Floor Discussion.

### Session 76: Genetics and Genomics: Methodology and Applications

Organizer: Ruzong Fan, Georgetown University Medical Center (GUMC).

Chair: Ruzong Fan, Georgetown University Medical Center (GUMC).

14:00 Ordered Multinomial Regression for Genetic Association Analysis of Ordinal Phenotypes at Biobank

*Jin Zhou*. University of Arizona

14:25 Gene—based association analysis of survival traits via functional regression—based mixed effect Cox models for related samples

♦ *Chi-Yang Chiu*<sup>1</sup> and *Ruzong Fan*<sup>2</sup>. <sup>1</sup>UTHSC <sup>2</sup>Georgetown University

14:50 Longitudinal Variant-Set Retrospective Association Test

♦ *Weimiao Wu and Zuoheng Wang*. Yale School of Public Health

15:15 Gene-based pleiotropic analysis of multiple survival traits via functional regressions with applications to eye diseases

*Bingsong Zhang*. Georgetown University

15:40 Floor Discussion.

### Session 74: Precision oncology trials: challenges and opportunities

Organizer: Weidong Zhang, Pfizer, Bo Huang, Pfizer.

Chair: Weidong Zhang, Pfizer.

14:00 On Design and Analysis of Biomarker-Integrated Clinical Trials with Adaptive Threshold Detection and Patient Enrichment

♦ *Xiaofei Wang*<sup>1</sup>, *Ting Wang*<sup>2</sup>, *Stephen George*<sup>1</sup> and *Haibo Zhou*<sup>2</sup>. <sup>1</sup>Duke University <sup>2</sup>Univ. of North Carolina at Chapel Hill

14:25 Group sequential enrichment designs based on adaptive regression of response and survival time on high dimensional covariates

♦ *Yeonhee Park*<sup>1</sup>, *Suyu Liu*<sup>2</sup>, *Peter Thall*<sup>2</sup> and *Ying Yuan*<sup>2</sup>. <sup>1</sup>Medical University of South Carolina <sup>2</sup>The University of Texas MD Anderson Cancer Center

14:50 Bayesian Semi-parametric Design (BSD) for Adaptive Dose-finding with Multiple Strata

♦ *Rachael Liu*<sup>1</sup>, *Jianchang Lin*<sup>1</sup>, *Mo Li*<sup>2</sup>, *Veronica Bunn*<sup>1</sup> and *Hongyu Zhao*<sup>2</sup>. <sup>1</sup>Takeda Pharmaceuticals <sup>2</sup>Yale University

15:15 Floor Discussion.

### Session 77: Applications of advanced statistics and artificial intelligence to genomics and precision medicine

Organizer: Momiao Xiong, University of Texas Health Science Center at Houston.

Chair: Shenying Fang, The University of Texas MD Anderson Cancer Center.

14:00 Recurrent Neural Reinforcement Learning for Counterfactual Evaluation of Public Health Interventions on the Spread of Covid-19 in the world

♦ *Qiyang Ge*<sup>1</sup>, *Zixin Hu*<sup>1</sup>, *Kai Zhang*<sup>2</sup>, *Tao Xu*<sup>2</sup>, *Shudi Li*<sup>2</sup>, *Wei Lin*<sup>1</sup>, *Li Jin*<sup>1</sup> and *Momiao Xiong*<sup>2</sup>. <sup>1</sup>Fudan University <sup>2</sup>The University of Texas Health Science Center at Houston

14:25 Combining Artificial Intelligence and Epidemiological Models for Prediction of COVID-19 In the US

♦ *Tao Xu*<sup>1</sup>, *Zhouxuan Li*<sup>1</sup>, *Kai Zhang*<sup>1</sup>, *Hongwen Deng*<sup>2</sup>, *Eric Boerwinkle*<sup>1</sup> and *Momiao Xiong*<sup>1</sup>. <sup>1</sup>The University of Texas Health Science Center at Houston <sup>2</sup>Tulane University

14:50 Floor Discussion.

### Session 75: Statistical Advancement and Challenges in Cell Therapy Development

Organizer: Rong Liu, Celgene.

Chair: Frank Shen, Celgene.

**Session 78: Innovative adaptive clinical trial designs**

Organizer: Yeting Du, Servier Pharmaceuticals.

Chair: Zhaoyang Teng, Servier Pharmaceuticals.

- 14:00 An optimal hybrid approach to calculate the conditional power  
*Jian Zhu*. Servier Pharmaceuticals
- 14:25 Innovative Adaptive Designs for Investigational Drug Development in Small Populations: A Discussion of Case Studies  
♦*Junjing Lin*<sup>1</sup>, *Godwin Yung*<sup>1</sup> and *Margaret Gamalo-Siebers*<sup>2</sup>. <sup>1</sup>Takeda Pharmaceuticals <sup>2</sup>Eli Lilly
- 14:50 ASIED: a Bayesian adaptive subgroup-identification enrichment design  
*Yanxun Xu*<sup>1</sup>, *Florica Constantine*<sup>1</sup>, ♦*Yuan Yuan*<sup>2</sup> and *Yili Pritchett*<sup>3</sup>. <sup>1</sup>Johns Hopkins University <sup>2</sup>AstraZeneca <sup>3</sup>Biometrics, G1 Therapeutics, Inc
- 15:15 Innovative Adaptive Designs for Investigational Drug Development in Small Populations: A Discussion of Case Studies  
*Godwin Yung*. Genentech
- 15:40 INSIGHt: A Bayesian Adaptive Platform Trial to Develop Precision Medicines for Patients With Glioblastoma  
*Lorenzo Trippa*. Dana-Farber Cancer Institute  
NA Floor Discussion.

**Session 79: Advanced Statistical Learning for High-dimensional Heterogeneous Data**

Organizer: Xiwei Tang, University of Virginia.

Chair: Xiwei Tang, University of Virginia.

- 14:00 Random projection pursuit regression  
*Qichen Liao*<sup>1</sup>, *Wei Zhang*<sup>1</sup>, *Jian Guo*<sup>2</sup> and ♦*Sijian Wang*<sup>3</sup>. <sup>1</sup>Tsinghua University <sup>2</sup>International Digital Economy Academy <sup>3</sup>Rutgers University
- 14:25 Subgroup Inference for Heterogeneous Treatment Effect Estimation  
♦*Lu Tang*<sup>1</sup> and *Ling Zhou*<sup>2</sup>. <sup>1</sup>University of Pittsburgh <sup>2</sup>Southwestern University of Finance and Economics
- 14:50 Joint Robust Multiple Inference on Large-Scale Multivariate Regression  
♦*Wen Zhou*<sup>1</sup>, *Youngseok Song*<sup>1</sup> and *Wenxin Zhou*<sup>2</sup>. <sup>1</sup>Colorado State University <sup>2</sup>University of California, San Diego
- 15:15 Multicategory Angle-based Learning for Estimating Optimal Dynamic Treatment Regimes with Censored Data  
♦*Fei Xue*<sup>1</sup>, *Yanqing Zhang*<sup>2</sup>, *Wenzhuo Zhou*<sup>3</sup>, *Haoda Fu*<sup>4</sup> and *Annie Qu*<sup>5</sup>. <sup>1</sup>University of Pennsylvania <sup>2</sup>Yunnan University <sup>3</sup>University of Illinois at Urbana-Champaign <sup>4</sup>Eli Lilly and Company <sup>5</sup>University of California Irvine
- 15:40 Floor Discussion.

**Session 80: New development in Bayesian methods and algorithms**

Organizer: Yang Ni, Texas A&amp;M University, Jianhua Huang, Texas A&amp;M University.

Chair: Yang Ni, Texas A&amp;M University.

- 14:00 Kriging: Beyond Matérn  
*Anindya Bhadra*. Purdue University
- 14:25 A Bayesian finite mixture model with variable selection for data with mixed-type variables  
♦*Shu Wang*<sup>1</sup> and *Chung-Chou Chang*<sup>2</sup>. <sup>1</sup>University of Florida <sup>2</sup>University of Pittsburgh
- 14:50 Ultra-Fast Approximate Inference Using Variational Functional Mixed Models  
*Shuning Huo*<sup>1</sup>, *Jeffrey S Morris*<sup>2</sup> and ♦*Hongxiao Zhu*<sup>1</sup>. <sup>1</sup>Virginia Tech <sup>2</sup>University of Pennsylvania
- 15:15 Recent Advances in Bayesian Methods for the Analysis of Tumor Pathology Images  
♦*Qiwei Li*<sup>1</sup>, *Cong Zhang*<sup>1</sup> and *Guanghua Xiao*<sup>2</sup>. <sup>1</sup>The University of Texas at Dallas <sup>2</sup>The University of Texas Southwestern Medical Center
- 15:40 Floor Discussion.

**Session 81: Innovative methods for complex censored data**

Organizer: Ruosha Li, The University of Texas Health Science Center at Houston.

Chair: Ying Ding, University of Pittsburgh.

- 14:00 Infinite Parameter Estimates in Proportional Hazards Regression  
*John Kolassa*<sup>1</sup> and ♦*Jane Zhang*<sup>2</sup>. <sup>1</sup>Rutgers University <sup>2</sup>Abbvie
- 14:25 Recent development on hierarchical joint frailty models for joint modeling of recurrent events and a terminal event  
*Zheng Li*<sup>1</sup>, *Vern M. Chinchilli*<sup>2</sup> and ♦*Ming Wang*<sup>2</sup>. <sup>1</sup>Novartis <sup>2</sup>Penn State College of Medicine
- 14:50 Conditional association and concordance for bivariate censored data  
*Ruosha Li*. The University of Texas Health Science Center at Houston
- 15:15 Regression with Covariates subject to Limits of Detection  
*Jimmy Kwon* and ♦*Bin Nan*. UC Irvine
- 15:40 Floor Discussion.

**Session 82: Student Paper Award Invited Session**

Organizer: Jing Ning, The University of Texas MD Anderson Cancer Center.

Chair: Jian Wang, The University of Texas MD Anderson Cancer Center.

- 14:00 BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data  
♦*Xinjun Wang*<sup>1</sup>, *Zhe Sun*<sup>1</sup>, *Yanfu Zhang*<sup>1</sup>, *Zhongli Xu*<sup>1</sup>, *Hongyi Xin*<sup>1</sup>, *Heng Huang*<sup>1</sup>, *Richard Duerr*, *Kong Chen*<sup>1</sup>, *Ying Ding*<sup>1</sup> and *Wei Chen*<sup>1</sup>. <sup>1</sup>University of Pittsburgh
- 14:20 Bayesian Meta-analysis of Censored Rare Events with Stochastic Coarsening  
♦*Xinyue Qi*<sup>1</sup>, *Christine B. Peterson*<sup>2</sup>, *Yucui Wang*<sup>3</sup> and *Shouhao Zhou*<sup>4</sup>. <sup>1</sup>The University of Texas Health Science Center at Houston <sup>2</sup>The University of Texas MD Anderson Cancer Center <sup>3</sup>Mayo Clinic <sup>4</sup>Pennsylvania State University
- 14:40 Inference for BART with Multinomial Outcomes  
♦*Yizhen Xu*<sup>1</sup> and *Joseph Hogan*<sup>2</sup>. <sup>1</sup>Johns Hopkins University <sup>2</sup>Brown University

15:00 Covariate Adaptive Family-wise Error Rate Control for Genome-Wide Association Studies  
♦ *Huijuan Zhou*<sup>1</sup>, *Xianyang Zhang*<sup>2</sup> and *Jun Chen*<sup>3</sup>.  
<sup>1</sup>Renmin University of China and Texas A&M University  
<sup>2</sup>Texas A&M University <sup>3</sup>Mayo Clinic

15:20 Functional Group Bridge for Simultaneous Regression and Support Estimation  
♦ *Zhengjia Wang*<sup>1</sup>, *John Magnotti*<sup>2</sup>, *Michael Beauchamp*<sup>2</sup> and *Meng Li*<sup>1</sup>. <sup>1</sup>Rice University <sup>2</sup>Baylor College of Medicine

15:40 Floor Discussion.

### Session 83: Statistical Methods for design and analysis of health outcomes

Organizer: Ling Chen, Washington University School of Medicine.  
Chair: Bin Zhang, University of Cincinnati.

14:00 The effect of biomarker variability on clinical outcomes: comparing different methods

♦ *Feng Gao*, *Jingqin Luo*, *Jinxia Liu*, *Chengjie Xiong* and *Lei Liu*. Washington University School of Medicine

14:25 Optimal designs in three-level cluster randomized trials with a binary outcome

♦ *Jingxia Liu*<sup>1</sup>, *Lei Liu*<sup>2</sup> and *Graham Colditz*<sup>1</sup>. <sup>1</sup>Washington University School of Medicine <sup>2</sup>Washington University School of Medicine (WUSM)

14:50 Matched or unmatched analyses with propensity-score-matched data?

*Fei Wan*. Washington university in St. Louis

15:15 Regression analysis of clustered failure time data with informative cluster size under the additive transformation models

♦ *Ling Chen*<sup>1</sup>, *Yanqin Feng*<sup>2</sup> and *Jianguo Sun*<sup>3</sup>.  
<sup>1</sup>Washington University in St Louis School of Medicine  
<sup>2</sup>Wuhan University <sup>3</sup>University of Missouri

15:40 Floor Discussion.

### Session 84: Statistical Modeling for COVID-19

Organizer: Yisheng Li, The University of Texas MD Anderson Cancer Center.

Chair: Jing Huang, The University of Pennsylvania.

14:00 Curating A COVID-19 Data Repository and Forecasting County-Level Death Counts in the United States

*Nick Altieri*<sup>1</sup>, *Rebecca L. Barter*<sup>1</sup>, *James Duncan*<sup>1</sup>, *Raaz Dwivedi*<sup>1</sup>, *Karl Kumbier*<sup>2</sup>, *Xiao Li*<sup>1</sup>, *Robert Netzorg*<sup>1</sup>, *Briton Park*<sup>1</sup>, *Chandan Singh*<sup>1</sup>, *Yan Shuo Tan*<sup>1</sup>, *Tiffany Tang*<sup>1</sup>, *Yu Wang*<sup>1</sup>, *Chao Zhang*<sup>1</sup> and ♦ *Bin Yu*<sup>1</sup>. <sup>1</sup>University of California, Berkeley <sup>2</sup>University of California, San Francisco

14:25 Semiparametric Bayesian Inference for the Transmission Dynamics of COVID-19 with a State-Space Model

*Tianjian Zhou*<sup>1</sup> and ♦ *Yuan Ji*<sup>2</sup>. <sup>1</sup>Colorado State University  
<sup>2</sup>The University of Chicago

14:50 Transmission dynamic modeling for COVID-19 data in U.S. and the world

♦ *Haoyu Zhang*<sup>1</sup>, *Chaolong Wang*<sup>2</sup> and *Xihong Lin*<sup>1</sup>.  
<sup>1</sup>Harvard T.H. Chan School of Public Health <sup>2</sup>Huazhong University of Science and Technology

15:15 A spatiotemporal model for county-level covid-19 infection data in the USA

*Peter Song*. University of Michigan

15:40 Floor Discussion.

### Dec. 15 16:00 - 17:40

#### Session 85: Recent statistical advancements in the design of clinical trials

Organizer: Tu Xu, Vertex.

Chair: Tu Xu, Vertex.

16:00 Innovative Group sequential Design with Optimal Futility Stopping rules

♦ *Zhaoyang Teng*<sup>1</sup>, *Qiang Zhao*<sup>2</sup>, *Rui Tang*<sup>1</sup> and *Yi Liu*<sup>2</sup>.  
<sup>1</sup>Servier Pharmaceuticals <sup>2</sup>Nektar Therapeutics

16:25 A Bayesian Adaptive Design for Concurrent Trials Involving Biologically-Related Diseases

♦ *Tony Jiang*<sup>1</sup>, *Amy Xia*<sup>1</sup>, *Matthew Psioda*<sup>2</sup>, *Joseph Ibrahim*<sup>2</sup> and *Jiawei Xu*<sup>2</sup>. <sup>1</sup>Amgen <sup>2</sup>UNC

16:50 A method for sample size calculation via E-value in the planning of observational studies

♦ *Yixin Fang*, *Weili He*, *Xiaofei Hu* and *Hongwei Wang*. Abbvie

17:15 Rethinking Treatment Switch in a randomized clinical trial (Innovative Design)

*Eiji Ishida*. FDA/CDER

17:40 Floor Discussion.

#### Session 86: Challenging Statistical Issues in Oncology Studies

Organizer: Jason Liao, Incyte.

Chair: Jing Yang, Merck.

16:00 Statistical Considerations on Using Minimal Residual Disease Status as the Efficacy Endpoint for Developing Novel Agents in Multiple Myeloma

♦ *Hong Tian*, *Jiajun Xu* and *Liang Xiu*. Johnson and Johnson

16:25 An adaptive seamless phase II/III design with simultaneous treatment and subpopulation selections in clinical trials with survival endpoints

*Cindy Lu*<sup>1</sup> and ♦ *Liwen Wu*<sup>2</sup>. <sup>1</sup>Biogen <sup>2</sup>University of Pittsburgh

16:50 Dynamic RMST Curves for Survival Analysis in Clinical Trials

♦ *Jason Liao*, *Frank Liu* and *Wen-Chi Wu*. Merck

17:15 Analysis of Time to Event Data using a Flexible Mixture Model under a Constraint of Proportional Hazards

♦ *Frank Liu*<sup>1</sup> and *Jason Liao*<sup>2</sup>. <sup>1</sup>Merck & Co., Inc. <sup>2</sup>Merck & Co., Inc.

17:40 Floor Discussion.

#### Session 87: Recent advances in machine learning and causal inference

Organizer: Yang Ning, Cornell University.

Chair: Yang Ning, Cornell University.

16:00 Sparsity Double Robust Inference of Average Treatment Effects  
*Jelena Bradic*<sup>1</sup>, *Stefan Wager*<sup>2</sup> and ♦*Yinchu Zhu*<sup>3</sup>. <sup>1</sup>UCSD  
<sup>2</sup>Stanford University <sup>3</sup>University of Oregon

16:25 Graph Quilting  
 ♦*Giuseppe Vinci*<sup>1</sup>, *Gautam Dasarathy*<sup>2</sup> and *Genevra Allen*<sup>3</sup>. <sup>1</sup>University of Notre Dame <sup>2</sup>Arizona State University <sup>3</sup>Rice University

16:50 Estimating the Effects of a New Technology using a Duration Model for Staggered Adoption (with Aureo de Paula (UCL))  
*Sida Peng*. Microsoft

17:15 Nonregular and Minimax Estimation of Individualized Thresholds in High dimension with Binary Responses  
*Yang Ning*. Cornell University

17:40 Floor Discussion.

### Session 88: Recent advances in accounting for heterogeneity in complex data

Organizer: Yong Chen, University of Pennsylvania.

Chair: Yong Chen, University of Pennsylvania.

16:00 A Statistical Framework for Genome-Scale Mutual Exclusivity Analysis of Cancer Mutations  
*Chi Wang*. University of Kentucky

16:25 Dissecting high-throughput (epi)genomics signals from heterogeneous samples  
*Hao Wu*. Emory University

16:50 Efficient Gene-Environment Interaction Tests for Large Biobank-Scale Sequencing Studies with Correlated Samples  
 ♦*Han Chen and Xinyu Wang*. The University of Texas Health Science Center at Houston

17:15 Modeling the dynamics of disease transmission  
 ♦*Jing Huang and Jiasheng Shi*. University of Pennsylvania

17:40 Floor Discussion.

### Session 89: Current advances in forensic statistics

Organizer: Larry Tang, University of Central Florida, Xiaochen Zhu, George Mason University.

Chair: Larry Tang, University of Central Florida.

16:00 Bayesian Characterizations Of U-processes Used In Pattern Recognition With Application To Forensic Source Identification  
 ♦*Christopher Saunders*<sup>1</sup>, *Cami Fuglsby*<sup>1</sup>, *Danica Ommen*<sup>2</sup> and *Joann Buscaglia*<sup>3</sup>. <sup>1</sup>South Dakota State University <sup>2</sup>Iowa State University <sup>3</sup>FBI Laboratory, Research and Support Unit

16:25 A Method of Forensic Evidence Interpretation using Error Rates  
 ♦*Danica Ommen*<sup>1</sup>, *Larry Tang*<sup>2</sup> and *Christopher Saunders*<sup>3</sup>. <sup>1</sup>Iowa State University <sup>2</sup>University of Central Florida <sup>3</sup>South Dakota State University

16:50 Univariate Likelihood Ratio Estimation via Mixture of Beta Distributions  
 ♦*Martin Slawski and He Qi*. George Mason University

17:15 Order-Constrained ROC Regression with Application to Facial Recognition

♦*Xiaochen Zhu*<sup>1</sup>, *Martin Slawski*<sup>2</sup> and *Larry Tang*<sup>3</sup>.  
<sup>1</sup>George Mason University <sup>2</sup>George Mason University  
<sup>3</sup>University of Central Florida

17:40 Floor Discussion.

### Session 90: Statistical and Machine Learning models on EHR and Insurance Claim databases

Organizer: Vahed Maroufy, The University of Texas Health Science Center at Houston.

Chair: Vahed Maroufy, The University of Texas Health Science Center at Houston.

16:00 Big Data to answer Big Questions: Experience with Anuerysmal SAH

*Vahed Mafoury*<sup>1</sup>, ♦*Ashraf Yaseen*<sup>1</sup> and *George Williams*<sup>2</sup>.  
<sup>1</sup>UTHSPA <sup>2</sup>McGovern Medical School

16:25 Statistics and Machine Learning Methods for EHR Data: From Data Extraction to Data Analytics/Predictions

*Ashraf Yaseen*. The University of Texas Health Science Center at Houston

16:50 Inferring Comorbidity Networks from EHR Data

♦*Xi Luo*<sup>1</sup>, *Gen Zhu*<sup>2</sup> and *Hulin Wu*<sup>2</sup>. <sup>1</sup>The University of Texas Health Science Center at Houston <sup>2</sup>University of Texas Health Science Center at Houston

17:15 Disease Influence Factor and Directed Disease Comorbidity Networks Derived from Big Longitudinal Health Care Data

*Vahed Maroufy*. The University of Texas Health Science Center at Houston

17:40 Floor Discussion.

### Session 91: Utilization of Historical Control Data for Clinical Development

Organizer: Jane Zhang, Allergan.

Chair: Jeen Liu, AbbVie.

16:00 To borrow or not to borrow? Determining when historical borrowing has value in a clinical trial

*Kert Viele*. Berry Consultants

16:25 Use of Pseudo Controls in Clinical Development

*Larry Shen*. Clinical Informatics-WuXi Apptec

16:50 Explore the use of matching method to supplement clinical trials with historical control data

*Xiang Zhang*. CSL Behring

17:15 Discussant

*Lu Tian*. Stanford University

17:40 Floor Discussion.

### Session 92: Statistical development for single-cell RNA-Seq data in biomedical studies

Organizer: Jianhua Hu, Columbia University.

Chair: Jianhua Hu, Columbia University.

16:00 scDesign2: a statistical simulator that recapitulates gene correlations for benchmarking scRNA-seq data analysis

♦*Tianyi Sun, Wei Vivian Li and Jingyi Jessica Li*. University of California, Los Angeles

- 16:25 Integrative differential expression and gene set enrichment analysis for scRNA-seq studies  
*Ying Ma<sup>1</sup>, Shiquan Sun<sup>1</sup>, Mengjie Chen<sup>2</sup> and ♦Xiang Zhou<sup>1</sup>.*  
<sup>1</sup>University of Michigan <sup>2</sup>University of Chicago
- 16:50 Imputation methods for scRNA-seq data  
 ♦*Wei Vivian Li<sup>1</sup> and Jingyi Jessica Li<sup>2</sup>.* <sup>1</sup>The State University of New Jersey <sup>2</sup>University of California, Los Angeles
- 17:15 RZiMM-scRNA: A regularized zero-inflated mixture model framework for single-cell RNA-seq data  
 ♦*Xinlei Mi and Jianhua Hu.* Columbia University
- 17:40 Floor Discussion.

17:15 Floor Discussion.

### Session 95: Real World Evidence Study in Healthcare

Organizer: Yahui Tian, Boehringer Ingelheim.

Chair: Yahui Tian, Boehringer Ingelheim.

- 16:00 Deep Learning for Analyzing Electronic Health Records  
*Fei Wang.* Weill Cornell Medicine
- 16:25 Statistical Anomaly Detection in Dynamic Brain Networks  
 ♦*Dorcas Ofori-Boateng<sup>1</sup>, Ivor Cribben<sup>2</sup> and Yulia Gel<sup>3</sup>.* <sup>1</sup>Portland State University <sup>2</sup>University of Alberta <sup>3</sup>University of Texas at Dallas
- 16:50 Disease screening for a personality disorder using Electronic Health Records (EHR) data  
 ♦*Nan Shao<sup>1</sup>, Marianne Goodman<sup>2</sup>, Chengxi Zang<sup>3</sup> and Vikas Mohan Sharma<sup>1</sup>.* <sup>1</sup>Boehringer Ingelheim <sup>2</sup>Icahn School of Medicine at Mount Sinai; James J Peters VA Medical Center <sup>3</sup>Weill Cornell Medicine, Cornell University
- 17:15 Estimation of Individualized Treatment Rules Using a Covariate-Specific Treatment Effect Curve  
 ♦*Wenchuan Guo<sup>1</sup>, Xiao-Hua Zhou<sup>2</sup> and Shujie Ma<sup>3</sup>.*  
<sup>1</sup>Bristol-Myers Squibb <sup>2</sup>Peking University <sup>3</sup>University of California Riverside
- 17:40 Floor Discussion.

### Session 93: Machine Learning and Real World Data

Organizer: Lan Zhou, Texas A&M University.

Chair: Lan Zhou, Texas A&M University.

- 16:00 New development of statistical machine learning methods with applications to large NHLBI longitudinal studies  
 ♦*Colin Wu<sup>1</sup>, Xiaoyang Ma<sup>1</sup> and Xin Tian<sup>2</sup>.* <sup>1</sup>National Heart, Lung and Blood Institute <sup>2</sup>National Heart, Lung and Blood Institute
- 16:25 A Practical Guideline for Factor Analysis of Binary/Ordinal Data  
 ♦*Wen Wan<sup>1</sup>, Vivian Li<sup>2</sup>, Erin Hoefling<sup>3</sup>, Dave Faldmo<sup>3</sup>, Rosy Chang Weir<sup>2</sup> and Marshall Chin<sup>1</sup>.* <sup>1</sup>University of Chicago <sup>2</sup>Association of Asian Pacific Community Health Organizations <sup>3</sup>Siouxland Community Health Center
- 16:50 Dynamic Risk Prediction Using Survival Tree Ensembles  
 ♦*Yifei Sun<sup>1</sup>, Sy Han Chiou<sup>2</sup>, Colin Wu<sup>3</sup>, Meghan McGarry<sup>4</sup> and Chiung-Yu Huang<sup>5</sup>.* <sup>1</sup>Columbia University <sup>2</sup>University of Texas at Dallas <sup>3</sup>National Heart, Lung, and Blood Institute <sup>4</sup>University of California San Francisco <sup>5</sup>University of California San Francisco
- 17:15 Harnessing Real-World Data for Regulatory Use and Applying Innovative Applications  
 ♦*Kelly Zou<sup>1</sup>, Jim Li<sup>1</sup>, Joseph Imperato<sup>1</sup>, Chandrashekhar Potkar<sup>1</sup>, Nikuj Sethi<sup>2</sup>, Jon Edwards<sup>3</sup> and Amrit Ray<sup>2</sup>.*  
<sup>1</sup>Viatrix <sup>2</sup>Pfizer <sup>3</sup>Envision Pharma Group
- 17:40 Floor Discussion.

### Session 96: Bayesian Additive Regression Tree: Theory, Computation, and Application

Organizer: Guanyu Hu, University of Connecticut.

Chair: Guanyu Hu, University of Connecticut.

- 16:00 Multidimensional Monotonicity Discovery with MBART  
 ♦*Robert McCulloch<sup>1</sup> and Edward George<sup>2</sup>.* <sup>1</sup>Arizona State University <sup>2</sup>University of Pennsylvania
- 16:25 Bayesian Decision Tree Ensembles in Fully Nonparametric Problems  
 ♦*Antonio Linero<sup>1</sup>, Yinpu Li<sup>2</sup> and Jared Murray<sup>1</sup>.*  
<sup>1</sup>University of Texas at Austin <sup>2</sup>Florida State University
- 16:50 Causal Inference and Sensitivity Analysis for Unmeasured Confounding in Observational Data with Multiple Treatments and a Binary Outcome  
 ♦*Liangyuan Hu<sup>1</sup>, Chenyang Gu<sup>2</sup>, Michael Lopez<sup>3</sup>, Jiayi Ji<sup>1</sup> and Juan Wisnivesky<sup>1</sup>.* <sup>1</sup>Icahn School of Medicine <sup>2</sup>Analysis Group, Inc. <sup>3</sup>Skidmore College
- 17:15 Stochastic tree ensembles for regularized nonlinear regression  
 ♦*Jingyu He<sup>1</sup> and P. Richard Hahn<sup>2</sup>.* <sup>1</sup>University of Chicago <sup>2</sup>Arizona State University
- 17:40 Floor Discussion.

### Session 94: Recent advances in statistical genomics, genetics and EHR data

Organizer: Yingying Wei, The Chinese University of Hong Kong.

Chair: Yingying Wei, Chinese University of Hong Kong.

- 16:00 WEVar: a novel statistical learning framework for predicting noncoding regulatory variants  
*Li Chen.* Indiana University School of Medicine
- 16:25 A Novel Approach on Multiple-Traits Genetic Association Tests for Flexible Pleiotropy Structures  
*Han Hao.* University of North Texas
- 16:50 Fused Landmark Approach for Dynamic Risk Prediction with Application to Electronic Health Record Data  
 ♦*Jiehuan Sun<sup>1</sup>, Katherine Liao<sup>2</sup> and Tianxi Cai<sup>3</sup>.*  
<sup>1</sup>University of Illinois at Chicago <sup>2</sup>Harvard Medical School <sup>3</sup>Harvard T.H. Chan School of Public Health

### Session 97: New Methods for Missing Data in Public Health Studies

Organizer: Jing Wu, University of Rhode Island.

Chair: Zhihua Ma, Shenzhen University.

- 16:00 An augmented survival analysis method for mis-measured and interval censored outcomes  
 ♦*Chongliang Luo, Rebecca Hubbard and Yong Chen.* University of Pennsylvania

- 16:25 Deep Learning for Time-to-event Outcomes  
♦Jon Steingrimsson, Samantha Morrison and Constantine Gatsonis. Brown University
- 16:50 A New Bayesian Joint Model for Mixed Types of Longitudinal Data in the Presence of Different Missing Data Patterns with Applications to HIV Prevention Trials  
Jing Wu. University of Rhode Island
- 17:15 Bayesian Modeling and Inference for Item Response Model with Nonignorable Missing Data  
Jing Wu<sup>1</sup>, ♦Zhihua Ma<sup>2</sup> and Ming-Hui Chen<sup>3</sup>. <sup>1</sup>University of Rhode Island <sup>2</sup>Shenzhen University <sup>3</sup>University of Connecticut
- 17:40 Floor Discussion.

**Dec. 16 9:00 - 10:00****Session 98: Keynote speech**

Organizer: Hulin Wu, University of Texas Health Science Center at Houston.

Chair: Hulin Wu, University of Texas Health Science Center at Houston.

- 9:00 Towards a Blend of Statistics and Microeconomics  
Michael I. Jordan. University of California, Berkeley

**Dec. 16 10:20 - 12:00****Session 99: Innovative Statistical and data science methods for clinical trial studies**

Organizer: Hongjian Zhu, The University of Texas Health Science Center at Houston.

Chair: Jun Yu, Abbvie.

- 10:20 Phase I/II seamless designs in oncology trials  
Inna Perevozskaya. GSK
- 10:45 Time to Endpoint Maturation Framework and Application  
♦Li Wang<sup>1</sup>, Mengjia Yu<sup>2</sup> and Hongtao Zhang<sup>3</sup>. <sup>1</sup>Abbvie <sup>2</sup>UIUC <sup>3</sup>Celgene
- 11:10 Response adaptive randomization designs and implementation in clinical trials  
Lanju Zhang. Abbvie
- 11:35 Precision medicine: subgroup identification in clinical trials  
♦Lei Liu and Chamila Perera. Washington University in St. Louis
- 12:00 Floor Discussion.

**Session 100: New methods in semiparametric inferences for analyzing real world data**

Organizer: Suojin Wang, Texas A&M University.

Chair: Lei Jin, Texas A&M University at Corpus Christi.

- 10:20 Empirical Likelihood for varying coefficient Geo Models  
Shuoyang Wang. Auburn University

- 10:45 Semi-parametric multinomial logistic regression for multivariate point pattern data  
Kristian Hesselund<sup>1</sup>, ♦Ganggang Xu<sup>2</sup>, Yongtao Guan<sup>2</sup> and Rasmus Waagepetersen<sup>1</sup>. <sup>1</sup>Aalborg University <sup>2</sup>University of Miami
- 11:10 Floor Discussion.

**Session 101: Recent development in Semiparametric regression analysis**

Organizer: Baojiang Chen, The University of Texas Health Science Center at Houston.

Chair: Baojiang Chen, The University of Texas Health Science Center at Houston.

- 10:20 Set regression with application in subgroup analysis  
Ao Yuan. Georgetown University
- 10:45 Semiparametric inference for marginal and association parameters in the distribution of bivariate event times data  
Dongdong Li<sup>1</sup>, Joan Hu<sup>2</sup> and ♦Rui Wang<sup>1</sup>. <sup>1</sup>Harvard Pilgrim Health Care Institute and Harvard Medical School <sup>2</sup>Simon Fraser University
- 11:10 Joint Penalized Spline Modeling of Multivariate Longitudinal Data  
♦Lihui Zhao<sup>1</sup>, Tom Chen<sup>2</sup>, Vladimir Novitsky<sup>2</sup> and Rui Wang<sup>2</sup>. <sup>1</sup>Northwestern University <sup>2</sup>Harvard University
- 11:35 Estimation and testing for extended partially linear single-index models  
Zijuan Chen and ♦Suojin Wang. Texas A&M University
- 12:00 Floor Discussion.

**Session 102: Stochastic gradient Monte Carlo for big data statistics**

Organizer: Faming Liang, Purdue University.

Chair: Faming Liang, Purdue University.

- 10:20 Extended Stochastic Gradient MCMC for Large-Scale Bayesian Variable Selection  
♦Qifan Song, Yan Sun, Mao Ye and Faming Liang. Purdue University
- 10:45 Optimal-Transport Bayesian Sampling  
Changyou Chen. University at Buffalo
- 11:10 Stochastic Gradient MCMC for Sequential Decision Making  
Yian Ma. UC San Diego
- 11:35 An Adaptively Weighted Stochastic Gradient MCMC Algorithm for Global Optimization in Deep Learning  
Wei Deng, Guang Lin and ♦Faming Liang. Purdue University
- 12:00 Floor Discussion.

**Session 103: Statistical and AI inferences based on DNA and protein sequences**

Organizer: Yun-Xin Fu, The University of Texas Health Science Center at Houston.

Chair: Xiaoming Liu, University of South Florida, Haipeng Li, Chinese Academy of Sciences.

10:20 Stairway Plot 2: demographic history inference with folded SNP frequency spectra  
 ♦ *Xiaoming Liu*<sup>1</sup> and *Yun-Xin Fu*<sup>2</sup>. <sup>1</sup>University of South Florida <sup>2</sup>The University of Texas Health Science Center at Houston

10:45 Origin of protein collective motions: a case study on serine protease proteinase K  
*Shu-Qun Liu*. Yunnan University

11:10 Statistical inferring the clonal and subclonal architecture of cancer genomes  
*Yupeng Cun*. Chinese Academy of Sciences

11:35 Supervised learning for analyzing large-scale genome-wide DNA polymorphism data  
*Haipeng Li*. Chinese Academy of Sciences

12:00 Floor Discussion.

11:35 Bayesian Additive Regression Trees for Causal Inference  
*Dai Feng*. AbbVie Inc.

12:00 Floor Discussion.

**Session 105: Xiangrong Yin Memorial Session**

Organizer: Wenbo Wu, The University of Texas at San Antonio.

Chair: Wenbo Wu, The University of Texas at San Antonio.

Speaker	Affiliation
Zhezhen Jin	Columbia University
Bing Li	Pennsylvania State University
Lexin Li	University of California at Berkeley
Yehua Li	University of California at Riverside
Hanxiang Peng	Indiana University-Purdue University Indianapolis
Xaofeng Shao	University of Illinois at Urbana-Champaign
T.N. Sriram	University of Georgia
John Stufken	University of North Carolina at Greensboro
Xiaogang Su	University of Texas at El Paso
Li Wang	Iowa State University
Qin Wang	University of Alabama
Yichao Wu	University of Illinois at Chicago
Yingcun Xia	National University of Singapore
Xin Zhang	Florida State University
Yichuan Zhao	Georgia State University

**Session 104: Advanced topics in causal inference**

Organizer: Meijing Wu, AbbVie Inc..

Chair: Tianshuang Wu, AbbVie Inc..

10:20 Variable Selection for Causal Mediation Analysis Using LASSO-based Methods  
*Zhaoxin Ye*<sup>1</sup>, ♦ *Yeying Zhu*<sup>1</sup> and *Donna Coffman*<sup>2</sup>.  
<sup>1</sup>University of Waterloo <sup>2</sup>Temple University

10:45 On regression approach to propensity score analysis  
*Liang Li*. The University of Texas MD Anderson Cancer Center

11:10 SSc/SSc-ILD Patient Journey: Data-driven Disease Trajectories in EHR/Claims Databases  
*Yahui Tian*. Boehringer Ingelheim

## Abstracts

### Session 1: Advancement of Machine Learning Methods via Tensors and High-Dimensional Tools

#### High-order Joint Embedding for Multi-Level Link Prediction

♦ *Yubai Yuan and Annie Qu*

University of California, Irvine  
yubaiy@uci.edu

Link prediction infers potential links from observed networks, and is one of the essential problems in network analyses. In contrast to traditional graph representation modeling which only predicts two-way pairwise relations, we propose a novel tensor-based joint network embedding approach on simultaneously encoding pairwise links and hyperlinks onto a latent space, which captures the dependency between pairwise and multi-way links in inferring potential unobserved hyperlinks. The major advantage of the proposed embedding procedure is that it incorporates both the pairwise relationships and subgroup-wise structure among nodes to capture richer network information. In addition, the proposed method introduces a hierarchical dependency among links to infer potential hyperlinks, and leads to better link prediction. In theory we establish the estimation consistency for the proposed embedding approach, and provide a faster convergence rate compared to link prediction utilizing pairwise links or hyperlinks only. Numerical studies on both simulation settings and Facebook ego-networks indicate that the proposed method improves both hyperlink and pairwise link prediction accuracy compared to existing link prediction algorithms.

#### Tensor denoising and completion based on ordinal observations

*Chanwoo Lee and ♦ Miaoyan Wang*

UW-Madison  
saplingwang@gmail.com

Higher-order tensors arise frequently in applications such as neuroimaging, recommendation system, social network analysis, and psychological studies. We consider the problem of low-rank tensor estimation from possibly incomplete, ordinal-valued observations. Two related problems are studied, one on tensor denoising and another on tensor completion. We propose a multi-linear cumulative link model, develop a rank-constrained M-estimator, and obtain theoretical accuracy guarantees. Our mean squared error bound enjoys a faster convergence rate than previous results, and we show that the proposed estimator is minimax optimal under the class of low-rank models. Furthermore, the procedure developed serves as an efficient completion method which guarantees consistent recovery of an order-K ( $d, \text{dots}, d$ )-dimensional low-rank tensor using only  $O(Kd)$  noisy, quantized observations. We demonstrate the outperformance of our approach over previous methods on the tasks of clustering and collaborative filtering.

#### Correlation Tensor Decomposition and Its Application in Spatial Imaging Data

*Yujia Deng<sup>1</sup>, ♦ Xiwei Tang<sup>2</sup> and Annie Qu<sup>3</sup>*

<sup>1</sup>University of Illinois Urbana-Champaign

<sup>2</sup>University of Virginia

<sup>3</sup>University of California Irvine  
xt4yj@virginia.edu

Multi-dimensional tensor data has gained increasing attention in recent years, especially in biomedical imaging analyses. However, most existing tensor models are only based on the mean information

of imaging pixels. Motivated by multimodal optical imaging data in a breast cancer study, we develop a new tensor learning approach to utilize pixel-wise correlation information, which is represented through the higher-order correlation tensor. We propose novel semi-symmetric correlation tensor decomposition method which effectively captures the informative spatial patterns of pixel-wise correlations to facilitate cancer diagnosis. We establish the theoretical properties for recovering structure and for classification consistency. In addition, we develop an efficient algorithm to achieve computational scalability. Our simulation studies and an application on breast cancer imaging data all indicate that the proposed method outperforms other competing methods in terms of pattern recognition and prediction accuracy.

#### Unconventional regression paradigm for microbiome compositional data with phylogenetic tree structure

*Gen Li*

University of Michigan  
ligen@umich.edu

Human microbiome is associated with many complex diseases. Microbiome data are usually represented as compositions, residing in a simplex that does not admit the standard Euclidean geometry. Moreover, there typically exists phylogenetic tree structure among microbial species which implies the evolutionary relationships among the microbes. Existing regression methods for microbiome data usually first apply some transformation to remove the compositional covariates from the simplex and then fit a standard regression model. However, the transformation and other preprocessing procedures (e.g., substituting zeros by some arbitrary small number) may induce bias and distort model interpretation. Moreover, it impedes the incorporation of the phylogenetic information. To address the issues, we develop a whole new regression paradigm for high dimensional compositional data with tree structure. We will directly model compositions as predictors and build an identifiable linear regression model. Special regularization based on the hierarchical tree structure will be imposed to improve model interpretation and taxonomic compatibility. Numerical studies show that the new paradigm can make full use of the phylogeny and turn the complex nature of microbiome data into a blessing.

### Session 2: Latest development in latent variable models and genetics

#### BUSseq: a Bayesian Hierarchical Model Providing One-stop Services for scRNA-seq Data

*Fangda Song, Ga Ming Angus Chan and ♦ Yingying Wei*

Chinese University of Hong Kong  
ywei@cuhk.edu.hk

There has been extensive research on computational methods for single-cell RNA-seq (scRNA-seq) Data. However, most existing methods are multi-stage approaches—clustering can only be performed after the batch effects have been corrected and the differential expressed genes can only be called after the cells have been clustered. The major issue with multi-stage methods is that uncertainties in the previous stages are often ignored. For instance, when cells have been first clustered into different cell types and then dif-



ferential gene expression identification is conducted, the clustering results are taken as if they were the underlying truth. As the clustering results may be prone to errors in practice, this can lead to false positives and false negatives. In contrast, we have recently developed Batch effects correction with Unknown Subtypes for scRNA-seq data (BUSseq), which is an interpretable Bayesian hierarchical model that can simultaneously correct batch effects, cluster cell types, impute missing data caused by dropout events, and detect differentially expressed genes without requiring a preliminary normalization step. Moreover, we mathematically show that in addition to the frequently advocated yet rarely implemented completely randomized experimental design, under two more flexible and realistic experimental designs—the reference panel and the chain-type designs—true biological variability can also be separated from batch effects for scRNA-seq data. We demonstrate that BUSseq outperforms existing methods with simulated and real data.

### Longitudinal Structural Topic Models for Estimating Latent Health Trajectories using Administrative Claims Data

Mengbing Li and ♦Zhenke Wu

University of Michigan, Ann Arbor  
zhenkewu@umich.edu

Administrative claims data present a unique opportunity for longitudinal assessment of patients' health. We adapt topic models, widely used for text mining, to analyzing such data. In this work, we estimate an unobserved patient-specific trajectory that characterizes her progression of multiple latent biological aberrations, each of which is an unobserved topic that yields distinct content distributions of the diagnosis codes. We propose a novel extension of the structural topic model (MargaretE. Roberts, 2016) that builds in important features of claims data: repeated multi-variate diagnosis codes, and time-varying covariates for topic prevalences and content distribution. Our model specifies the topic prevalences by logistic mixed models and the content distributions by regularized logistic models. We derive a Gibbs sampler for inference using P olya-Gamma augmentation. We apply the model to data from 15K cancer-associated thrombosis patients extracted from OptumInsight claims database. By aggregating monthly diagnosis codes (ICD-9) over multiple months as correlated documents from a patient, we quantify the latent disease progression and the effects of baseline and time-varying covariates.

### Latent variable modeling in biomarker studies

Zheyu Wang

Johns Hopkins University  
wangzy@jhu.edu

Accumulating evidence suggest that the initiation of the Alzheimer disease (AD) pathogenic process precede the first symptoms by a decade or more. The recognition of this decade-long asymptomatic stage has greatly impact AD research and therapeutic development to focus on preclinical stage of AD pathogenic process, at which time disease modifying therapy is more likely to be effective. On the other hand, the decade-long preclinical stage imposes a major challenge in investigating biomarkers for early AD detection, because 1) using clinical diagnosis as the reference point can be in error, especially in the early course of the disease; and 2) most AD studies do not have autopsy data to confirm diagnoses. Until technology advance allows for brain examination with "autopsy level" clarity, an appropriate statistical method that directly address the unobservable nature of preclinical AD progression is necessary for any rigorous AD biomarker evaluation and for efficient analyzing AD study data where only clinical data are available and neuropathology data are

not yet available. Since AD pathophysiology has been recognized as a multidimensional process that involves amyloid deposition, neurofibrillary tangles and neurodegeneration among other aspects, we propose latent variable model to study the underlying AD pathophysiology process revealed by multidimensional markers and apply the model to two different AD data sets.

### Accounting for correlated horizontal pleiotropy in two-sample Mendelian randomization using correlated instrumental variants

Qing Cheng and ♦Jin Liu

Duke-NUS Medical School  
jin.liu@duke-nus.edu.sg

Mendelian randomization (MR) is a powerful approach to examine the causal relationships between health risk factors and outcomes from observational studies. Due to the proliferation of genome-wide association studies (GWASs) and abundant fully accessible GWASs summary statistics, a variety of two-sample MR methods for summary data have been developed to either detect or account for horizontal pleiotropy, primarily based on the assumption that the effects of variants on exposure ( $\gamma$ ) and horizontal pleiotropy ( $\alpha$ ) are independent. This assumption is too strict and can be easily violated because of the correlated horizontal pleiotropy (CHP). To account for this CHP, we propose a Bayesian approach, MR-Corr2, that uses the orthogonal projection to reparameterize the bivariate normal distribution for  $\gamma$  and  $\alpha$ , and a spike-slab prior to mitigate the impact of CHP. We develop an efficient algorithm with paralleled Gibbs sampling. To demonstrate the advantages of MR-Corr2 over existing methods, we conducted comprehensive simulation studies to compare for both type-I error control and point estimates in various scenarios. By applying MR-Corr2 to study the relationships between pairs in two sets of complex traits, we did not identify the contradictory causal relationship between HDL-c and CAD. Moreover, the results provide a new perspective of the causal network among complex traits.

### Session 3: New methods for joint analysis of survival and longitudinal data

#### Joint Analysis of Interval Censored Survival Time and Longitudinal Data

Di Wu and ♦Chenxi Li

Michigan State University  
cli@msu.edu

Joint analysis of time-to-event and longitudinal data is needed to fit survival models with intermittently observed time-dependent covariates. The literature is scarce in joint analysis of longitudinal data and interval censored failure time, especially for multivariate interval censored survival data. We develop a joint analysis method for longitudinal data and (multivariate) interval censored failure time to fit a class of semiparametric (mixed effects) transformation models with a longitudinal covariate under interval censoring. The finite sample performance of the method is evaluated in simulations. An application to a dental caries data set illustrate its utility.

#### Semiparametric latent-class models for multivariate longitudinal and survival data

♦Kin Yau Wong<sup>1</sup>, Donglin Zeng<sup>2</sup> and Dan-Yu Lin<sup>2</sup>

<sup>1</sup>Hong Kong Polytechnic University

<sup>2</sup>University of North Carolina at Chapel Hill

kin-yau.wong@polyu.edu.hk

In long-term follow-up studies, data are often collected on repeated measures of multivariate response variables as well as on time to the occurrence of a certain event. To jointly analyze such longitudinal data and survival time, we propose a general class of semiparametric latent-class models that accommodates a heterogeneous study population with flexible dependence structures between the longitudinal and survival outcomes. We combine nonparametric maximum likelihood estimation with sieve estimation and devise an efficient EM algorithm to implement the proposed approach. We establish the asymptotic properties of the proposed estimators through novel use of modern empirical process theory, sieve estimation theory, and semiparametric efficiency theory. Finally, we demonstrate the advantages of the proposed methods through extensive simulation studies and provide an application to a major epidemiological cohort study.

#### **Regression analysis with proportional intensity model for general mixed recurrent event data with terminal events**

♦ *Liang Zhu*<sup>1</sup>, *Yimei Li*<sup>2</sup> and *Gregory T. Armstrong*<sup>3</sup>

<sup>1</sup>The University of Texas Health Science Center at Houston

<sup>2</sup>St. Jude Children's Research Hospital

<sup>3</sup>St. Jude Children's Research Hospital  
liang.zhu@uth.tmc.edu

Recurrent events data occur frequently in longitudinal studies across numerous fields of biomedical research. For example, they comprise a key component of the information collected in the largest childhood cancer survivorship study in North America. Because recording recurrent events requires recording a composite set of event times and numbers, the completeness of recurrent events data collected for each individual subject may vary. In the statistical literature, the different levels of completeness are often labeled as recurrent event, panel count, panel binary, or panel ordinal data. Recurrent event data record the occurrence time of each event, while panel count data record the count of events, panel ordinal data record a category to which the count of event belongs, and panel binary data record whether any event has occurred during a given observation period. Among the four data types, recurrent event data is the hardest to collect but contains the most information, followed by panel count data, and then panel ordinal and panel binary data. Many studies contain a mix of these different data types due to the ever-present challenges associated with data collection - for example, missing follow-up data or inconsistent questionnaire design. We propose a proportional intensity model that is able to integrate all available information, regardless of the data type, for the regression analysis of the general mixed recurrent event data. Moreover, we consider the complications caused by a terminal event such as death, which can also be related to the underlying recurrent event process. We propose joint frailty models to account for the correlation between the recurrent event process and the terminal event. We further establish estimation procedures for our proposals. Simulation studies demonstrate that the methods work well in realistic situations. We then apply the proposed method to analyze hospitalization rates in childhood cancer survivors.

#### **Cost-effective analysis for active surveillance versus nephron-sparing surgery for Bosniak III renal cysts: An application of multistate model**

*Xu Zhang*

The University of Texas Health Science Center at Houston  
xu.zhang@uth.tmc.edu

Nephron-sparing surgery (NSS) is the standard treatment for patients with Bosniak III renal cysts. We conducted a cost-

effectiveness study to evaluate active surveillance (AS) versus NSS. We developed the multistate models to depict possible transitions for a patient with Bosniak III renal cysts undergoing AS or NSS. Based on published results of similar studies, we estimated transition intensities as well as costs of procedures and utilities associated with different health states. Utilizing the product-limit estimators and utilities, we obtained the estimates of transition probabilities and quality-adjusted lifetimes (QALY). Finally we evaluated the incremental cost-effective ratio of NSS relative to AS for 60-year-old man and woman. Our base-case estimates suggested that AS yielded longer QALY and lower lifetime cost compared to NSS.

#### **Session 4: The Data are BIG and We are PRECISE: New Statistical Methods for Precision Medicine.**

##### **Bayesian nonparametric survival regression for optimizing precision dosing of intravenous busulfan in allogeneic stem cell transplantation**

*Peter Thall*

The University of Texas MD Anderson Cancer Center  
rex@mdanderson.org

Intravenous busulfan has become a standard component of the preparative regimen in allogeneic stem cell transplantation (alloSCT) for acute leukemia. Systemic busulfan exposure, characterized by the area under the plasma concentration versus time curve, AUC, is strongly associated with clinical outcome. An AUC that is too high is associated with severe toxicities, while an AUC that is too low carries increased risks of disease recurrence and graft failure. Consequently, an optimal AUC interval must be determined for each patient by giving a preclinical dose. To address the possibility that the optimal AUC interval may vary with individual patient characteristics, we developed a tailored approach for determining optimal covariate-specific AUC-intervals. To estimate personalized AUC intervals, we applied a flexible Bayesian non-parametric survival regression model based on a dependent Dirichlet process and Gaussian process prior (DDP-GP) to analyze historical data from 151 alloSCT patients. The fitted model identified optimal AUC intervals that varied substantively with age and whether the patient was in complete remission or had active disease at transplant. Potentially, these results may change clinical practice in alloSCT worldwide. Extensive simulations showed that the DDP-GP regression model's performance compares favorably with several alternative robust methods. An R package, DDPGPSurv, that implements the DDP-GP model for a broad range of survival regression analyses is provided.

##### **Inferring Longitudinal antiretroviral drugs effects on depressive symptomatology in homogenous people with HIV**

♦ *Yanxun Xu*<sup>1</sup>, *Wei Jin*<sup>1</sup>, *Yang Ni*<sup>2</sup> and *Leah Rubin*<sup>1</sup>

<sup>1</sup>Johns Hopkins University

<sup>2</sup>Texas A&M University  
yanxun.xu@jhu.edu

The effects of antiretroviral (ART) drugs for people living with HIV (PLWH) on depressive symptomatology are inconsistent. Given the heterogeneous nature of both ART drugs and the presentation of depressive symptoms, newer approaches are necessary for guiding clinical practice. Since ART-related depression would be heterogeneous among HIV patients depending on their differences in numerous factors including demographics and clinical variables, we develop a new Bayesian semiparametric graphical model with nodes representing drugs and depression items, and weighted edges rep-

resenting their relationships. The weights indicate the strength of the drug-depression relationships and can vary across different visits and different patients. The effective and reliable modeling and prediction will help elucidate the treatment-depression relationship and guide the clinicians in making more informed decision for patients.

### **Precision medicine for patients with chronic liver diseases through medical imaging**

*Peng Zhang*

University of Michigan  
pczhang@med.umich.edu

In this talk, we present our approach to extract high-dimensional phenotyping of morphological data from medical imaging based on machine learning and deep learning methods. These novel risk markers can be utilized to provide precision medicine for an array of outcomes for patients with chronic liver disease, including cirrhosis, mortality, hepatocellular carcinoma.

### **Kernel-Involved-Dosage-Decision Learning method for estimating dynamic dosage regimes**

♦*Ming Tang*<sup>1</sup>, *Matthew Schipper*<sup>2</sup>, *Theodore Lawrence*<sup>2</sup> and *Lu Wang*<sup>2</sup>

<sup>1</sup>Boehringer Ingelheim (China) investment Co. Ltd

<sup>2</sup>University of Michigan  
mingtang@umich.edu

Dose-finding plays a critical role in medical research and drug development. When treating patients with chronic disease, clinicians need to adjust the treatment dose over time to accommodate patients' disease progression. Therefore, a sequence of dose assignments, one per stage, is chosen by physicians by accounting for the observed time-varying medical characteristics and other baseline histories. This paradigm of routinely treating patients is known as a dynamic treatment regimes (DTRs). In the case of continuous dosage, a DTR (also known as dynamic dosage regime) consists of decision stages, one per stage, mapping individualized patient characteristics to a dose assignment. Due to the complexity of the continuous dose scale, few of the existing literature has developed methods for estimating the optimal dynamic dosage regime. We propose a personalized dose-finding method, kernel-involved-dosage-decision learning (KIDD-Learning) method, which combines a robust estimation of the dose-response curve with an interpretable decision tree for estimating the optimal dynamic dosage regime in a multiple-stage setting. At each stage, KIDD-Learning recursively estimates the robust dose-response function using kernel regression and then grow a decision tree to estimate the optimal dosage for each stage. Simulation studies demonstrate the performance of KIDD-Learning under different scenarios. We also apply KIDD-Learning to Michigan Medicine Liver SBRT data to evaluate the dose assignments of adaptive radiation therapy and estimate the optimal dynamic dosage regime for patients with liver cancer.

## **Session 5: Decipher cell heterogeneity in high-throughput data analysis**

### **Choice of Scale in the Estimation of Cell-type Proportions**

♦*Johann Gagnon-Bartsch*<sup>1</sup> and *Gregory Hunt*<sup>2</sup>

<sup>1</sup>University of Michigan

<sup>2</sup>College of William and Mary  
johanngb@umich.edu

Complex tissues are composed of a large number of different types of cells, each involved in a multitude of biological processes. Con-

sequently, an important component to understanding such processes is understanding the cell-type composition of the tissues. Estimating cell type composition using high-throughput gene expression data is known as cell-type deconvolution. In this talk, we first summarize the extensive deconvolution literature by identifying a common regression-like approach to deconvolution. We call this approach the Unified Deconvolution-as-Regression (UDAR) framework. While methods that fall under this framework all use a similar model, they fit using data on different scales. Two popular scales for gene expression data are logarithmic and linear. Unfortunately, each of these scales has problems in the UDAR framework. Using log-scale gene expressions proposes a biologically implausible model and using linear-scale gene expressions will lead to statistically inefficient estimators. To overcome these problems, we propose a new approach for cell-type deconvolution that works on a hybrid of the two scales. This new approach is biologically plausible and improves statistical efficiency. We compare the hybrid approach to other methods on simulations as well as a collection of eleven real benchmark datasets. Here, we find the hybrid approach to be accurate and robust.

### **In silico cell type deconvolution by integrative single-cell RNA-seq and bulk RNA-seq analysis**

*Mingyao Li*

University of Pennsylvania  
mingyao@pennmedicine.upenn.edu

Knowledge of cell type composition in disease relevant tissues is an important step towards the identification of cellular targets of disease. In this talk, I will present a method that utilizes cell-type specific gene expression from single-cell RNA-seq data to characterize cell type compositions from bulk RNA-seq data in complex tissues from diverse samples. By iteratively identifying cell type invariant genes between disease conditions and appropriately weighting of genes showing cross-subject and cross-cell consistency, our method can transfer cell type-specific gene expression information from one dataset to another, and infer cell type compositions in diverse samples. We further show that the estimated cell type proportions allow us to characterize allele-specific expression (ASE) with cell type resolution in bulk RNA-seq data. To do so, we regress the bulk level allele-specific read counts over estimated cell-type proportions through a linear mixed-effect model, and test for the presence of ASE in each cell type. Extensive evaluations show that this method is powerful in detecting cell type-specific ASE effect even for rare cell types.

### **Tumor cell total mRNA expression shapes the molecular and clinical phenotype of cancer**

*Shaolong Cao*<sup>1</sup>, *Jennifer R. Wang*<sup>1</sup>, *Shuangxi Ji*<sup>1</sup>, *Peng Yang*<sup>1</sup>, *Matthew D. Montierth*<sup>1</sup>, *Shuai Guo*<sup>1</sup>, *John Paul Shen*<sup>1</sup>, *Xiao Zhao*<sup>1</sup>, *Jingxiao Chen*<sup>1</sup>, *Alfonso Urbanucci*<sup>2</sup>, *Jonas Demeulemeester*<sup>3</sup>, *Peter Van Loo*<sup>3</sup> and ♦*Wenyi Wang*<sup>1</sup>

<sup>1</sup>The University of Texas MD Anderson Cancer Center

<sup>2</sup>Oslo University Hospital

<sup>3</sup>The Francis Crick Institute  
wwang7@mdanderson.org

Cancers can vary greatly in their transcriptomes. In contrast to alterations in specific genes or pathways, differences in tumor cell total mRNA content has not been comprehensively assessed. Technical and analytical challenges have impeded examination of total mRNA expression at scale across cancers. To address this, we evaluated total mRNA expression using single cell sequencing, and developed a model for quantifying tumor-specific total mRNA expression (TmS)

from bulk sequencing data. We estimated and validated TmS in 5,205 patients across 15 cancer types identifying significant inter-individual variability. At a pan-cancer level, high TmS is associated with increased risk of disease progression and death. Cancer type-specific patterns of genetic alterations, intra-tumor heterogeneity, as well as pan-cancer trends in metabolic dysregulation and hypoxia contribute to TmS. Taken together, our results suggest that measuring total mRNA expression offers a broader perspective of tracking cancer transcriptomes, which has important clinical and biological implications.

#### Cell type-specific Expression Quantitative Trait Loci

*Little Paul<sup>1</sup>, Dan-Yu Lin<sup>2</sup>, Yun Li<sup>2</sup> and Wei Sun<sup>1</sup>*

<sup>1</sup>Fred Hutchinson Cancer Research Center

<sup>2</sup>University of North Carolina at Chapel Hill

wsun@bios.unc.edu

Gene expression vary across cell types, which at least partially explain the diverse morphologies and phenotypes across cell types despite almost identical DNA sequences. Genetic regulation also vary across cell types and traditional gene expression quantitative trait locus (eQTL) analyses often ignore such heterogeneity and map eQTLs of bulk samples as if they are composed of homogeneous cells. We have developed a new method to perform cell type-specific eQTL mapping for bulk RNA-seq data. Our method utilize both total expression and allele-specific expression, which further improves the power of eQTL mapping. We evaluated our method using various simulations and gene expression from GTEx whole blood samples, GTEx brain samples, as well as CommonMind Consortium (CMC) brain samples.

### Session 6: Recent Development on Heterogeneity Analysis

#### Integrated quantile rank test for gene-level associations in sequencing studies

*Tianying Wang, Iuliana Ionita-Laza and Ying Wei*

Columbia University

tianyingw0905@outlook.com

Testing gene-based associations is the fundamental approach to identify genetic associations in sequencing studies. It is also commonly used in other genetic association studies as an effective yet biologically meaningful way to enhance the statistical power. The best-known approaches include Burden and Sequence Kernel Association Tests (SKAT). The gene-traits associations are often complex due to population heterogeneity, gene-environmental interactions, and various other reasons. The mean-based tests, including Burden and SKAT, may miss or underestimate some high-order associations that could be scientifically interesting. In this paper, we propose a new family of gene-level association tests, which integrate quantile rank score processes while combining multiple weighting schemes to accommodate complex associations. The resulting test statistics enjoy multiple advantages. They are as efficient as the mean-based SKAT and Burden test when the associations are homogeneous across quantile levels and have improved efficiency for complex and heterogeneous associations. The test statistics are distribution-free, and could hence accommodate a wide range of distributions. They are also computationally feasible. We established the asymptotic properties of the proposed tests under the null and alternative hypothesis and conducted large scale simulation studies to investigate its finite sample performance. We applied the proposed tests to Metabochip data to identify genetic associations

with lipid traits and compared the results with those of the Burden and SKAT tests.

#### Regression Trees for Interval-Censored Data

*Ce Yang, Liqun Diao and Richard Cook*

University of Waterloo

anthree@gmail.com

When a failure process is under an intermittent observation scheme the failure status is only available at periodic assessment times resulting in interval-censored data. Despite the presence of interval censoring in a training sample, interest often lies in prediction based on such data. We consider the use of censoring unbiased transformations and pseudo-observations to define observed data loss functions which are unbiased estimates of complete data loss functions, and we use these to fit regression trees and make predictions using interval-censored data. The trees grown based on these methods are found have good properties empirically in terms of recovery of the tree structure and prediction accuracy. An application is given to a study of involving individuals with psoriatic arthritis where the aim is to identify genetic markers associated with the development of the axial disease.

#### Simultaneous prediction intervals for high-dimensional vector autoregressive model

*Mengyu Xu*

University of Central Florida

mengyu.xu@ucf.edu

The simultaneous prediction intervals for high-dimensional vector autoregressive model are studied. We consider a de-biased calibration for the lasso prediction and propose a Gaussian-multiplier bootstrap based method for one-step-ahead prediction. The asymptotic coverage consistency of the prediction interval is obtained. We also develop simulation results to evaluate the finite sample performance of the procedure.

### Session 7: Current Development in Experimental Designs and Its Applications

#### On design orthogonality, projection uniformity and maximin distance for computer experiments

*Yaping Wang<sup>1</sup>, Fasheng Sun<sup>2</sup> and Hongquan Xu<sup>3</sup>*

<sup>1</sup>East China Normal University

<sup>2</sup>Northeast China Normal University

<sup>3</sup>UCLA

ypwang@fem.ecnu.edu.cn

Column-orthogonality and maximin distance are two popular criteria for space-filling designs in both computer and physical experiments. Uniform projection designs concerning the design projection properties were recently proposed and showed to be strongly consistent to maximin L1-equidistant designs. In this paper we further investigate the connections among the criteria-column-orthogonality, projection uniformity and maximin (L1 and L2) distance. We show that the average squared correlation metric is a function of the pairwise L2-distances between the rows only. Based on this relationship we develop some new optimality and near optimality results and show that the three different criteria are closely related in some general cases. A new lower bound of the uniform projection criterion and two constructions of optimal designs under all of these criteria are also presented. These results not only provide new theoretical justifications for each criterion but also help in finding better space-filling designs.

**A method of constructing maximin distance designs**♦Wenlong Li<sup>1</sup>, Min-Qian Liu<sup>1</sup> and Boxin Tang<sup>2</sup><sup>1</sup>Nankai University<sup>2</sup>Simon Fraser University

wenlongli@mail.nankai.edu.cn

One attractive class of space-filling designs for computer experiments is that of maximin distance designs. Algorithmic search for such designs is commonly used but this method becomes ineffective for large problems. Theoretical construction of maximin distance designs is challenging; some results have been obtained recently, often by employing highly specialized techniques. This paper presents an easy-to-use method for constructing maximin distance designs. The method is versatile as it is applicable for any distance measure. Our basic idea is to construct large designs from small designs and the method is effective because the quality of large designs is guaranteed by that of small designs, as evaluated by the maximin distance criterion.

**Sequential good lattice point sets for computer experiments**♦Xueru Zhang<sup>1</sup>, Yong-Dao Zhou<sup>1</sup>, Dennis Lin<sup>2</sup> and Min-Qian Liu<sup>1</sup><sup>1</sup>Nankai university<sup>2</sup>Purdue University

zhangxueru2019@gmail.com

Sequential Latin hypercube designs (SLHDs) have recently received great attention for computer experiments. Existing approaches for constructing SLHDs are limited with respect to the sizes of run and factors and are infeasible for high dimensions with large run sizes. To overcome these challenges, we propose a new class of SLHDs called the sequential good lattice point (SGLP) sets. These SGLP sets can be applied to a wide variety of experimental spaces, such as the invariant space, the progressively contracting space and the mixed space which keeps either invariant or contracting in each sequential stage, with the flexibility in the run size and the number of factors. Moreover, we introduce the approaches for constructing space-filling SGLP (SSGLP) sets under a specified criterion that are fast and efficient, for cases with high dimensions and/or large run sizes. Furthermore, it is shown that the SSGLP set has a better space-filling property than the existing SLHDs in the invariant space. Consequently, we develop a local search method based on SGLP sets for solving non-differential optimization problem.

**Session 8: Keynote speech****Learning from COVID-19 Data in Wuhan, USA and the World on Transmission, Health Outcomes and Interventions**

Xihong Lin

Harvard University

xlin@hsph.harvard.edu

COVID-19 is an emerging respiratory infectious disease that has become a pandemic. In this talk, I will first provide a historical overview of the epidemic in Wuhan. I will provide the analysis results of 32,000 lab-confirmed COVID-19 cases in Wuhan to estimate transmission rates, the multi-faceted public health intervention effects that helped Wuhan control the COVID-19 outbreak, and epidemiological characteristics of the cases. I will present the results using the transmission dynamic model that show two features of the COVID-19 epidemic: high transmissibility and high covertness, and a high proportion of undetected cases, including asymptomatic and mildly symptomatic cases, and the chances of resurgence in different scenarios. I will next present the epidemic models to estimate the transmission rates in USA and other countries and intervention effects, as well as the prevalence and the total number of infections.

I will present methods and analysis results of > 500,000 participants of the HowWeFeel project on symptoms and health conditions in US, and discuss the factors associated with who have been tested in US and the factors associated with positive PRC tests/COVID-19 infection. I will provide several takeaways learned from the pandemic and discuss priorities.

**Session 9: Bayesian Methodology and Applications for Complex Biomedical Data****High dimensional mediation model for neuroimaging data analysis**Xiaoqing Wang<sup>1</sup>, Yimei Li<sup>1</sup>, Wilburn Reddick<sup>1</sup>, Heather Conklin<sup>1</sup>, Amar Gajjar<sup>1</sup>, Cheng Cheng<sup>1</sup> and ♦Zhaohua Lu<sup>1</sup>St. Jude Children's Research Hospital

zhaohua.lu@stjude.org

Treatments for pediatric cancer have been reported to cause damage to the brain microstructure and undermine neurocognitive development. In this study, we propose a high dimensional mediation model to validate the hypothesis that part or all of the treatment impact on neurocognitive development is mediated by the brain microstructure. Bayesian spatial variable selection prior was used to identify the important brain regions that mediate the treatment impact on the neurocognitive impairment. Simulation studies were used to demonstrate the performance of the proposed method and compare it with some existing neuroimaging analysis methods.

**Bayesian inferences for panel count data and interval-censored data with nonparametric modeling of the baseline functions**♦Lu Wang<sup>1</sup>, Xiaoyan Lin<sup>2</sup> and Lianming Wang<sup>2</sup><sup>1</sup>Western New England University<sup>2</sup>University of South Carolina

lu.wang@wne.edu

Both panel-count data and interval-censored data arise commonly when individuals in a study are examined at periodic follow-ups. Interval-censored data are studied when the exact times of the events are of interest and these exact times are not directly observed but are only known to fall within some intervals formed by the observation times. Panel count data are under investigation when the exact times of the recurrent events are not of interest but the counts of the recurrent events occurring within the time intervals are available and of interest. A novel unified Bayesian approach is developed for analyzing panel count data under the Gamma frailty Poisson process model and interval-censored data under Cox's proportional hazards model and the proportional odds model. The baseline functions in those models share the same property of being nondecreasing positive functions and are modeled nonparametrically by assigning a Gamma process prior. Efficient and easy-to-implement Gibbs samplers are developed for the posterior computation under these three models for the two types of data. The proposed methods are evaluated in extensive simulation studies and illustrated by real-life data applications.

**Bayesian Latent Factor on Image Regression with Nonignorable Missing Data**

Xiaoqing (Jade) Wang

St. Jude Children's Research Hospital

jadexqwang@gmail.com

Medical imaging data have been widely used in modern health care, particularly in the prognosis, screening, diagnosis, and treatment of various diseases. In this study, we consider a latent factor-on-image (LoI) regression model that regresses a latent factor on ul-

trahigh dimensional imaging covariates. The latent factor is characterized by multiple manifest variables through a factor analysis model, while the manifest variables are subject to non-ignorable missingness. We propose a two-stage approach for statistical inference. At the first stage, an efficient functional principal component analysis method is applied to reduce the dimension and extract useful features/eigenimages. At the second stage, a factor analysis mode is proposed to characterize the latent response variable. Moreover, an LoI model is used to detect influential risk factors, and an exponential tiling model applied to accommodate nonignorable nonresponses. A fully Bayesian method with an adjust spike-and-slab absolute shrinkage and selection operator (lasso) procedure is developed for the estimation and selection of influential features/eigenimages. Simulation studies show the proposed method exhibits satisfactory performance. The proposed methodology is applied to a study on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset.

### Bayesian Semiparametric Regression Analysis of Multivariate Panel Count Data

Chunling Wang<sup>1</sup> and Xiaoyan Lin<sup>2</sup>

<sup>1</sup>University of South Carolian

<sup>2</sup>University of South Carolina  
lin9@mailbox.sc.edu

Panel count data often occur in a long-term recurrent event study, where the exact occurrence times of the recurrent events are unknown, but only the occurrence counts between any two adjacent observation time points are recorded. Most traditional methods only handle panel count data for a single type of event. In this paper, we propose a Bayesian semiparametric approach to analyze panel count data for multiple types of events. For each type of event, the proportional mean model is adopted to model the mean count of the event, where its baseline mean function is approximated by monotone I-splines (Ramsay 1988). The correlation between multiple events is modeled by common frailty terms and scale parameters. Unlike many frequentist estimating equation methods, our approach is based on the observed likelihood and makes no assumption on the relationship between the recurrent processes and the observation process. Under the Poisson process assumption, we develop an efficient Gibbs sampler based on a novel data augmentation for the MCMC sampling. Simulation studies show good estimation performance of the baseline mean functions and the regression coefficients; meanwhile the importance of including the scale parameter to flexibly accommodate the correlation between events is also demonstrated. Finally, a skin cancer data example is fully analyzed to illustrate the proposed methods.

### Session 10: New Challenges in Lifetime Data Analyses

#### Analysis of the Time-Varying Cox Model for Cause-Specific Hazard Functions With Missing Causes

Fei Heng<sup>1</sup>, Yanqing Sun<sup>2</sup>, Seunggeun Hyun<sup>3</sup> and Peter Gilbert<sup>4</sup>

<sup>1</sup>University of North Florida

<sup>2</sup>University of North Carolina at Charlotte

<sup>3</sup>University of South Carolina Upstate

<sup>4</sup>Fred Hutchinson Cancer Research Center  
f.heng@unf.edu

This paper studies the Cox model with time-varying coefficients for cause-specific hazard functions when the causes of failure are subject to missingness. Inverse probability weighted and augmented inverse probability weighted estimators are investigated. The latter is considered as a two-stage estimator by directly utilizing the inverse

probability weighted estimator and through modeling available auxiliary variables to improve efficiency. The asymptotic properties of the two estimators are investigated. Hypothesis testing procedures are developed to test the null hypotheses that the covariate effects are zero and that the covariate effects are constant. Simulation studies are conducted to examine the finite sample properties of the proposed estimation, and hypothesis testing procedures under various settings of the auxiliary variables and the percentages of the failure causes that are missing. It demonstrates that the augmented inverse probability weighted estimators are more efficient than the inverse probability weighted estimators, and that the proposed testing procedures have the expected satisfactory results in sizes and powers. The proposed methods are illustrated using the Mashi clinical trial data for investigating the effect of randomization to formula-feeding versus breastfeeding plus extended infant zidovudine prophylaxis on death due to mother-to-child HIV transmission in Botswana.

### Regression Analysis of Mixed Panel Count Data with Dependent Terminal Events

Guanglei Yu<sup>1</sup>, Liang Zhu<sup>2</sup>, Yang Li<sup>3</sup>, Jianguo Sun<sup>4</sup> and Leslie Robison<sup>5</sup>

<sup>1</sup>Eli Lilly and Company

<sup>2</sup>The University of Texas Health Science Center at Houston

<sup>3</sup>University of North Carolina at Charlotte

<sup>4</sup>University of Missouri-Columbia

<sup>5</sup>St. Jude Children's Research Hospital  
yli18@iu.edu

Event history studies are commonly conducted in many fields and a great deal of literature has been established for the analysis of the two types of data commonly arising from these studies: recurrent event data and panel count data. The former arises if all study subjects are followed continuously, while the latter means that each study subject is observed only at discrete time points. In reality, a third type of data, a mixture of the two types of the data above, may occur and furthermore, as with the first two types of the data, there may exist a dependent terminal event, which may preclude the occurrences of recurrent events of interest. This talk discusses regression analysis of mixed recurrent event and panel count data in the presence of a terminal event and an estimating equation-based approach is proposed for estimation of regression parameters of interest. In addition, the asymptotic properties of the proposed estimator are established and a simulation study conducted to assess the finite-sample performance of the proposed method suggests that it works well in practical situations. Finally the methodology is applied to a childhood cancer study that motivated this study.

### Semiparametric Estimation of the Cure Fraction in Population-based Cancer Survival Analysis

Ennan Gu<sup>1</sup>, Jiajia Zhang<sup>1</sup>, Wenbin Lu<sup>2</sup>, Lianming Wang<sup>3</sup> and Federico Felizzi<sup>4</sup>

<sup>1</sup>University of South Carolina

<sup>2</sup>North Carolina State University

<sup>3</sup>University of South Carolina, SC

<sup>4</sup>F. Hoffmann-La Roche Ltd, Basel  
jzhang@mailbox.sc.edu

With rapid development in medical research, the treatment of diseases including cancer has progressed dramatically and those survivors may die from causes other than the one under study, especially among elderly patients. Motivated by the SEER female breast cancer study, background mortality is incorporated into the mixture cure proportional hazards (MCPH) model to improve the cure fraction estimation in population-based cancer studies. Here, that pa-

tients are “cured” is defined as when the mortality rate of the individuals in diseased group returns to the same level as that expected in the general population, where the population level mortality is presented by the mortality table of the United States. The semiparametric estimation method based on the EM algorithm for the MCPH model with background mortality (MCPH+BM) is further developed and validated via comprehensive simulation studies. Real data analysis shows that the proposed semiparametric MCPH+BM model may provide more accurate estimation in population-level cancer study.

### **Benefit-harm Tradeoff in Individualized Treatment with Censored Data**

*Shuai Chen*

University of California, Davis  
shschen@ucdavis.edu

It is widely recognized that treatments often have substantially different effects across the population. Many statistical methods have recently been developed for identifying subgroups of patients who may benefit from different available treatments. It is important to investigate individualized benefit-harm tradeoff regarding treatment strategies (e.g., treatment may be toxic). In many clinical studies to evaluate the treatment benefits (e.g., survival time) and harms (e.g., toxicity duration or medical costs), censored data pose challenges to the analysis. Due to the induced dependent censoring problem, standard survival analysis techniques are often invalid for censored costs and toxicity duration. We propose a method for estimating individualized treatment benefits and harms with censored data in both randomized clinical trials and observational studies, which would provide a tool for physicians and patients to make decision based on personalized benefit-harm tradeoff. Our method bypasses the modelling of main effect, and hence involves minimum modeling for the relationship between the outcome and covariates pertinent to measuring individual treatment benefit-harm tradeoff. The proposed method also allows variable selection via regularization. We then conducted numerical studies to evaluate the performance of proposed method.

### **Session 11: Novel Semiparametric and Machine Learning Tools in Complex Observational Studies**

#### **Case-cohort Studies with Multiple Interval-censored Disease Outcomes**

♦ *Qingning Zhou<sup>1</sup>, Jianwen Cai<sup>2</sup> and Haibo Zhou<sup>2</sup>*

<sup>1</sup>University of North Carolina at Charlotte

<sup>2</sup>University of North Carolina at Chapel Hill  
qzhou8@uncc.edu

Interval-censored failure time data commonly arise in epidemiological and biomedical studies where the occurrence of an event or a disease is determined via periodic examinations. Subject to interval-censoring, available information on the failure time can be quite limited. Cost-effective sampling designs are desirable to enhance the study power, especially when the disease rate is low and the covariates are expensive to obtain. In this work, we formulate the case-cohort design with multiple interval-censored disease outcomes and generalize it to nonrare diseases where only a portion of diseased subjects are sampled. We develop a marginal sieve weighted likelihood approach, which assumes that the failure times marginally follow the proportional hazards model. We consider two types of weights to account for the sampling bias, and adopt a sieve method with Bernstein polynomials to handle the unknown baseline

functions. We employ a weighted bootstrap procedure to obtain a variance estimate that is robust to the dependence structure between failure times. The proposed method is examined via simulations and illustrated with the ARIC data.

#### **svReg: Structural Varying-coefficient Regression Identifies Individualized Relationship between Brain Regions and Motor Impairment in Huntington Disease**

♦ *Rakheon Kim<sup>1</sup>, Samuel Mueller<sup>2</sup> and Tanya Garcia<sup>1</sup>*

<sup>1</sup>Texas A&M University

<sup>2</sup>University of Sydney

rkim@stat.tamu.edu

For Huntington disease, developing interventions which target individual patients or specific groups of patients requires identification of brain regions related to motor impairment. Statistically, this can be cast as a varying-coefficient model selection but it is challenging when the coefficients are modified by structured variables such as categorical variables. We propose a novel variable selection method to account for these structured modifying variables. Our method is empirically shown to better screen irrelevant categorical modifying variables than existing methods which do not account for the structured modifying variables. Also, our method screens irrelevant variables better than the existing methods even for continuous modifying variables and main predictors. Hence, our method leads to a model with lower false discovery rate and higher prediction accuracy than the existing methods. Finally, we show the relationship between brain regions and motor impairment is different for each subgroup of the patients depending on their disease severity. To the best of our knowledge, our study is the first to identify such interaction effects between the disease severity and brain regions, which indicates the need for personalized intervention by disease severity.

#### **Efficient Semiparametric Inference for Two-Phase Studies with Outcome and Covariate Measurement Errors**

*Ran Tao*

Vanderbilt University Medical Center

r.tao@vumc.org

In modern observational studies using electronic health records or other routinely collected data, both the outcome and covariates of interest can be error-prone and their errors often correlated. A cost-effective solution is the two-phase design, under which the error-prone outcome and covariates are observed for all subjects during the first phase and that information is used to select a validation subsample for accurate measurements of these variables in the second phase. Previous research on two-phase measurement error problems largely focused on scenarios where there are errors in covariates only or the validation sample is a simple random sample of study subjects. Herein, we propose a semiparametric approach to general two-phase measurement error problems with a quantitative outcome, allowing for correlated errors in the outcome and covariates and arbitrary second-phase selection. We devise a computationally efficient and numerically stable expectation-maximization algorithm to maximize the nonparametric likelihood function. The resulting estimators possess desired statistical properties. We demonstrate the superiority of the proposed methods over existing approaches through extensive simulation studies, and we illustrate their use in an observational HIV study.

## Session 12: Statistical Inference and Modeling for High-Dimensional and Complex Data Structure

### The Conditional Adaptive Lasso and Its Sufficient Variable Selection

Chenlu Ke

Virginia Commonwealth University  
kec2@vcu.edu

Lasso and Adaptive Lasso have been popular and powerful techniques for variable selection. When some prior knowledge shows importance for a certain set of variables, then given this set to select additional relevant variables becomes important for building better models. This leads to the Conditional Adaptive Lasso. We show that the Conditional Adaptive Lasso enjoys the oracle properties; namely, it performs as well as if the underlying model were given in advance. We provide Sufficient Conditional Adaptive Lasso algorithms for both coefficient estimation and variable selection. Furthermore, we extend the Conditional Adaptive Lasso to adopt to the generalized linear models, i.e. Generalized Conditional Adaptive Lasso with coefficient estimation and variable selection. Numerical studies of comparing with penalized approaches and other screening methods are conducted to demonstrate the advantages of our methods.

### Specification tests for covariance structures in high-dimensional statistical models

Xiao Guo<sup>1</sup> and Cheng Yong Tang<sup>2</sup>

<sup>1</sup>University of Science and Technology of China

<sup>2</sup>Temple University  
yongtang@temple.edu

We consider testing the specifications of the covariance structures in statistical models for high-dimensional data. In particular, we are interested in developing such tests when the random vectors of interests are not directly observable, and have to be derived from some estimated models. Additionally, the covariance specifications may involve extra nuisance parameters whose estimations are required. In a generic additive model setting, we develop and investigate test statistics based on the maximum discrepancy measure calculated from the residuals. To approximate the distributions of the test statistics under the null hypothesis, new multiplier bootstrap procedures are proposed with necessary adjustments incorporating the model and nuisance parameter estimation errors. Our theoretical development elucidates the impact due to the model and parameter estimation errors in different settings, and establishes the validity of our testing procedures. Extensive simulations and real data examples confirm the results from our analysis, and demonstrate the performance of the specification tests.

### Subspace Estimation with Automatic Dimension and Variable Selection in Sufficient Dimension Reduction

Jing Zeng, Qing Mai and Xin Zhang

Florida State University  
mai@stat.fsu.edu

Sufficient dimension reduction (SDR) methods target at finding lower-dimensional representations of a multivariate predictor  $X$  such that all the information about the conditional distribution of the response  $Y$  given  $X$  is preserved. The reduction is commonly achieved by projecting the predictor onto a low-dimensional subspace. The smallest such subspace is known as the Central Subspace (CS), and is the key parameter of interest for most SDR methods. In this article, we propose a unified and flexible framework for estimating the CS in high dimensions. Our approach generalizes a wide range of model-based and model-free SDR methods

to high-dimensional settings, where the CS is assumed to involve only a subset of the predictors for interpretability. We formulate the problem as a quadratic convex optimization so that the global solution is feasible. The proposed estimation procedure simultaneously achieves the structural dimension selection and coordinate-independent variable selection of the CS. Theoretically, our method achieves both dimension selection and subspace estimation consistency under mild conditions. We also demonstrate the effectiveness and efficiency of our method with extensive simulation studies and real data examples.

### Pseudo Estimation in Regression

Wenbo Wu<sup>1</sup> and Xiangrong Yin<sup>2</sup>

<sup>1</sup>University of Texas at San Antonio

<sup>2</sup>University of Kentucky  
wenbo.wu@utsa.edu

Often a problem in linear regression either for correlated data or high-dimensional data is the singularity of the sample covariance matrix. While dealing with an ill-conditioned covariance matrix has been a longstanding challenge in the statistical literature, recent proposals relying on adding noises to the original data have found their successes in obtaining reliable estimates and making predictions. It has been shown that perturbing data with noises has a close relationship to the well-known ridge estimator. In this paper, we propose to add noises to the predictors with a known covariance structure, and call the estimator obtained in such a way a “pseudo estimator”. A new variable selection procedure based on the concept of pseudo confidence interval which works for both correlated predictors and “large  $p$  small  $n$ ” problem is proposed. We study the theoretical properties of the proposed pseudo estimator and variable selection procedure. Furthermore, we use an ensemble step to stabilize the pseudo estimation (variable selection) results. The advantages of the proposed method are demonstrated by both simulation studies and real data analyses.

## Session 13: Artificial Intelligence and Causal Inference

### Neural causal network learning

Momiao Xiong, Tao Xu and Yuanyuan Liu

The University of Texas Health Science Center at Houston  
momiao.xiong@uth.tmc.edu

Learning causal networks is a daunting task due to both the combinatorial feature of the structure searching space and causal identifiability. The causal networks often take the form of a directed acyclic graph (DAG). Learning DAGs is a NP hard problem. There are two basic types of methods for causal structure learning: Constrained-based (testing conditional independence) and score-based methods<sup>30</sup>. The main stream of causal structure learning is the score-based methods. The score-based methods formulate the causal learning problems in terms of optimizing a certain score (objective) function with unknown adjacency matrix and constraints that ensure that the graph is acyclic. The past optimization problems of the score-based methods have combinatorial nature and hence are difficult to solve. Recently, some researchers propose to formulate the original combinatorial optimization problems of searching optimal causal networks into a continuous constrained optimization problem and cleverly use the matrix of exponential of the adjacency matrix to pose constraints of the acyclicity of the graph. This approach has four remarkable features. The first feature is that this approach uses neural DAG learning. This will provide a powerful tool for nonlinear causal inference. The second feature is that similar to deep learning, we can use gradient-based optimization methods



to solve the reformulated continuous optimization problems of the causal learning. The third feature is that the acyclicity constraints are both efficiently computable and easily differentiable. The fourth feature is that each node variable (e.g. gene) is a vector of features than can include gene expression, genomic information and epigenomic information for each gene. This provides a general framework for construction omics networks using all omics information. Neural DAG learning will open a new way for learning omics causal networks.

#### **Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection**

*Qiyang Ge<sup>1</sup>, Xuelin Huang<sup>2</sup>, Shenying Fang<sup>2</sup>, Shicheng Guo<sup>3</sup>, Wei Lin<sup>4</sup> and Momiao Xiong<sup>1</sup>*

<sup>1</sup>The University of Texas Health Science Center at Houston

<sup>2</sup>The University of Texas MD Anderson Cancer Center

<sup>3</sup>University of Wisconsin-Madison

<sup>4</sup>Fudan University

xlhuang@mdanderson.org

Artificial intelligence (AI) is a powerful tool for precision oncology. It can accurately estimate the individualized treatment effects and learn optimal treatment choices. Therefore, the AI approach can substantially improve progress and treatment outcomes of patients. As one of AI approaches, conditional generative adversarial nets for inference of individualized treatment effects (GANITE) have been developed. However, the GANITE can only deal with binary treatment and does not provide a tool for optimal treatment selection. To overcome these limitations, we modify conditional generative adversarial networks (MCGANs) to allow estimation of individualized effects of any types of treatments including binary, categorical and continuous treatments. We propose to use sparse techniques for selection of biomarkers that predict the best treatment for each patient.

#### **Conditional generative adversarial networks and variational autoencoders for individualized biomarker selection and treatment effect estimation**

*Shenying Fang<sup>2</sup>, Qiyang Ge<sup>1,2</sup>, Shicheng Guo<sup>3</sup>, Yuanyuan Liu<sup>1</sup>, Jeffrey E. Lee<sup>2</sup>, Wei Lin<sup>4</sup> and Momiao Xiong<sup>1</sup>*

<sup>1</sup>The University of Texas Health Science Center at Houston

<sup>2</sup>The University of Texas MD Anderson Cancer Center

<sup>3</sup>University of Wisconsin-Madison

<sup>4</sup>Fudan University

sfang71@gmail.com

Next generation genomic, epigenomic, sensing and image technologies produce ever deeper multiple omics, physiological, imaging and phenotypic data with millions of features. Integrating omics and physiological data provides invaluable information for identification of individualized biomarkers that will be used for estimation of individualized treatment effects and optimal selection of individualized therapy. The classical methods for biomarker identification use average treatment effect information. However, treatment response is heterogeneous. Only using average treatment effect information presents a problem in selecting the optimal treatment for each individual to ensure that the right therapy is offered to “The right patient at the right time.” Unfortunately, estimating individualized treatment effects and design of individualized therapy is beyond the state-of-the-art of the current biomarker selection paradigm. A key issue for individualized treatment estimation is to estimate counterfactuals of treatment. However, counterfactuals are unobserved and are therefore a missing value problem. The classical treatment effect estimation methods cannot accurately estimate the counterfactuals

due to lack of methods for missing value estimation. The recently developed conditional generative adversarial nets (CGAN) are accurate tools for estimating the counterfactuals. Omics and physiological data involve millions of features. Since neural networks (NN) are complicated nonlinear functions, identifying biomarkers from omics data in CGAN is a challenging task. The variational autoencoder (VAE) is used to reduce the data dimension. To evaluate the performance of the CGAN approach to treatment effect estimation compared to the state-of-art methods, we first ran simulation study, by generating 4,189 treated and 4,111 untreated subjects, respectively. Then the VAE-based CGANs were applied to real TCGA dataset. The dataset included 18 cancer types and 2,196 subjects (787 radiotherapy only, 761 chemotherapy only and 648 both radiotherapy and chemotherapy) with both profiles of 1,881 miRNAs, and age, sex and cancer type information. Simulation and Real data analysis results show that the proposed algorithms substantially outperformed the state-of-the-art methods. Our results suggest that our new approach performs well to estimate individualized treatment effect and guide treatment selection.

#### **Artificial Intelligence and Causal Inference Inspired Methods for Forecasting the Spread of Covid-19 in the United States**

*Zixin Hu<sup>1</sup>, Qiyang Ge<sup>1</sup>, Shudi Li<sup>2</sup>, Eric Boerwinkle<sup>2</sup>, Wei Li<sup>1</sup>, Li Jin<sup>1</sup> and Momiao Xiong<sup>2</sup>*

<sup>1</sup>Fudan University

<sup>2</sup>The University of Texas Health Science Center at Houston

hu.zixin@foxmail.com

As of October 14, 2020, the number of cumulative cases of COVID-19 in the US exceeded 7,860,281 and included 215,955 deaths, thus causing a serious public health crisis. Meanwhile, an economic crisis and resistance to the strict intervention measures are rising. Some researchers proposed intermittent social distancing that may drive the outbreak of Covid-19 into 2022. Questions arise about whether we should maintain or relax or reverse quarantine measures at this time. We developed novel artificial intelligence and causal inference integrated methods for real-time identification, prediction and control of nonlinear time-varying epidemic dynamic systems. We estimated the peak time of the Covid-19, the peak number of cumulative cases in the US and when its outbreak in the US will be over if the current intervention measures remain in place. We also evaluated the impact of relaxing the current interventions for reopening economy on the spread of Covid-19. We provide analytic tools for balancing the health risks of workers and consumers, and reopening economy.

#### **Session 14: New advances in modern statistical modeling and testing**

##### **Spatiotemporal Autoregressive Partially Linear Varying Coefficient Models**

*Shan Yu, Lily Wang and Lei Gao*

Iowa State University

lilywang@iastate.edu

With growingly abundant data that relate to both space and time becoming available, spatiotemporal modeling has received increasing attention in the econometric literature. This paper targets on developing a class of spatiotemporal autoregressive partially linear varying coefficient models that are sufficiently flexible to simultaneously capture the spatiotemporal dependence and nonstationarity often encountered in practice. When spatial observations are observed over time and exhibit dynamic and nonstationary behaviors,

our models become a useful tool for analyzing such data. We develop a numerically stable and computationally efficient estimation procedure using the tensor product splines over triangular prisms to approximate the coefficient functions. The estimators of both the constant coefficients and varying coefficients are consistent. We also show that the estimators of the constant coefficients are asymptotically normal, which enables us to construct confidence intervals and make inferences. The performance of the method is evaluated by Monte Carlo experiments and applied to the analysis a house price data in Sydney.

#### **A new bootstrap assisted stationarity test in the time domain**

◆ *Lei Jin<sup>1</sup> and Suojin Wang<sup>2</sup>*

<sup>1</sup>Texas A&M Corpus Christi

<sup>2</sup>Texas A&M

lei.jin@tamucc.edu

Econometric modeling via time series analysis is vital in finance, with applications to the prediction of interest rates, foreign currency risk, stock market volatility, and so on. The stationarity is a critical assumption for the dynamic modeling in economics. While many previous stationarity tests have been developed for linear or Gaussian time series, time series are often nonlinear and non-Gaussian in many econometrics applications. In this paper, a bootstrap assisted test is proposed to check the stationarity of nonlinear time series. The test is based on a Walsh transformation in a framework of nonlinear time series by generalizing the stationarity test in Jin et al (2015). The asymptotic normality of the Walsh ordinates and their asymptotic covariance matrix under the null hypothesis are derived. The new form of the asymptotic covariance matrix is adaptive to nonlinearity, which is then consistently estimated by a bootstrap procedure. A simulation study is conducted to examine the finite sample performance of the test with comparisons to some existing competing methods, indicating that the proposed approach works well for nonlinear time series. The proposed test is applied to an analysis of a financial data set.

#### **Comparison of Difference Based Variance Estimators for Partially Linear Models**

◆ *Guoyi Zhang and Yan Lu*

University of New Mexico

gzhang@unm.edu

Recently, there has been increasing interest and activity in the general area of partially linear models, in which a parametric component is used to model the group effect, and a nonparametric component is used to model an underlying function that represents some certain shared environment. In this research, we evaluated two difference based variance estimators: Gasser, Sroka, and Jennen-Steinmetz (1986) (GSJS) and Hall, Kay, and Titterton (1990) (HKT) for use in partially linear models. Under various settings, we compared power of tests for heteroskedasticity, and other finite population properties of the estimators.

### **Session 15: Robust Methods in Missing Data and Causal Inference**

#### **Pattern graphs: a graphical approach to nonmonotone missing data problems**

◆ *Yen-Chi Chen and Mauricio Sadinle*

University of Washington

yenchi@uw.edu

In this paper, we introduce the concept of pattern graphs—a directed acyclic graph representing how response patterns are associated.

Pattern graphs provide an elegant way to model nonmonotone missing data. We introduce a selection model and a pattern mixture model formulation using the pattern graphs and show that they are equivalent. Pattern graphs lead to an inverse probability weighting estimator as well as an imputation-based estimator for estimating a parameter of interest. Asymptotic theories of the estimators are studied and we provide a graph-based dynamic programming procedure for computing both estimators. We introduce three graph-based sensitivity analysis and study the equivalence class under a generalized version of pattern graphs.

#### **The Promises of Parallel Outcomes**

*Ying Zhou, Dehan Kong and ◆ Linbo Wang*

University of Toronto

linbo.wang@utoronto.ca

Unobserved confounding presents a major threat to the validity of causal inference from observational studies. In this paper, we introduce a novel framework that leverages the information in multiple parallel outcomes for identification and estimation of causal effects. Under a shared confounding structure among multiple parallel outcomes, we achieve nonparametric identification with at least three parallel outcomes. We further show that under a set of linear structural equation models, causal inference is possible with two parallel outcomes. We also develop accompanying estimating procedures and evaluate their finite sample performance through simulation studies and a data application studying the causal effect of the tau protein level on various types of behavioral deficits.

#### **Propensity Score Calibration with Missing-at-Random Data**

*Peisong Han*

University of Michigan

peisong@umich.edu

Methods for propensity score (PS) calibration are commonly used in missing data analysis. Most of them are derived based on constrained optimizations where the form of calibration is dictated by the objective function being optimized and the calibration variables used in the constraints. Considerable efforts on pairing an appropriate objective function with the calibration constraints are usually needed to achieve certain efficiency and robustness properties for the final estimators. We consider an alternative approach where the calibration is carried out by solving the empirical version of certain moment equalities. Based on this approach, under the setting of estimating the mean of a response, we study how to achieve intrinsic efficiency and multiple robustness in the presence of multiple data distribution models.

#### **CCmed: Cross-condition mediation analysis for identifying robust trans-associations mediated by cis-gene**

◆ *Fan Yang<sup>1</sup>, Kevin Gleason<sup>2</sup>, Jiebiao Wang<sup>3</sup>, Jubao Duan<sup>4</sup>, Xin He<sup>2</sup>, Brandon Pierce<sup>2</sup> and Lin Chen*

<sup>1</sup>University of Colorado Anschutz Medical Campus

<sup>2</sup>University of Chicago

<sup>3</sup>University of Pittsburgh

<sup>4</sup>NorthShore University Health System

fan.3.yang@cuanschutz.edu

Trans-acting expression quantitative trait loci (eQTLs) collectively explain a substantial proportion of expression variation, yet are challenging to detect and replicate since their effects are often individually weak. A large proportion of genetic effects on distal genes are mediated through cis-gene expression. Cis-association (between SNP and cis-gene) and gene-gene correlation conditional on SNP genotype could establish trans-association (between SNP and trans-gene). Since both cis-association and gene-gene conditional corre-

lation tend to have effects shared across tissue types and conditions, trans-associations mediated by cis-expression often have robust effects shared across conditions and studies. We proposed a Cross-Condition Mediation analysis method - CCmed - to detect cis-mediated trans-associations by integrating cis-association and gene-gene conditional correlation statistics from multiple studies, allowing study heterogeneity. CCmed serves as a complementary strategy to the standard trans-association tests. We analyzed data from 13 brain tissue types from the Genotype-Tissue Expression (GTEx) project, and identified trios with cis-mediated trans-associations, many of which show evidence of replication in other data sets. We also identified trans-genes associated with known schizophrenia susceptibility loci and further validated the risk-associations of certain identified trans-genes by harnessing GWAS summary statistics from the Psychiatric Genomics Consortium and eQTL statistics from GTEx.

## Session 16: Functional Data Analysis: Theory and Application

### Optimal Function-on-Function Regression with Interaction between Functional Predictors

Honghe Jin<sup>1</sup>, ♦Xiaoxiao Sun<sup>2</sup> and Pang Du<sup>3</sup>

<sup>1</sup>University of Georgia

<sup>2</sup>University of Arizona

<sup>3</sup>Virginia Tech

xiaosun@email.arizona.edu

We consider a functional regression model in the framework of reproducing kernel Hilbert space where the interaction effect of two functional predictors, as well as their main effects, over the functional response is of interest. Compared to the traditional function-on-function regression, we only include one trivariate coefficient function in our model. The component functions for the main and interaction effects can be obtained through the functional ANOVA decomposition of this trivariate coefficient function. We show that our trivariate estimator achieves the minimax convergence rate in terms of mean prediction under the reproducing kernel Hilbert space framework. Furthermore, we propose an estimation procedure that can be easily implemented via standard numerical tools. Extensive numerical studies demonstrate the advantages of the proposed method over existing ones in terms of prediction and estimation of coefficient functions.

### Stochastic Functional Linear Models and Malliavin Calculus

♦Jin Zhou<sup>1</sup>, Weimiao Wu<sup>2</sup>, Chi-Yang Chiu<sup>3</sup> and Bingsong Zhang<sup>4</sup>

<sup>1</sup>University of Arizona

<sup>2</sup>Yale University

<sup>3</sup>University of Tennessee, Health Science Center

<sup>4</sup>Georgetown University Medical Center

rf740@georgetown.edu

We study stochastic functional linear models (SFLM) driven by an underlying square integrable stochastic process  $X(t)$  which is generated by a standard Brownian motion. Utilizing the magnificent Ito integrals and Malliavin calculus,  $X(t)$  is expanded into a summation of orthogonal multiple integrals. Based on the expansion, we show that the fourth moments of linear functionals of underlying stochastic process  $X(t)$  are bounded by the square of their second moments when  $X(t)$  is a finite linear combination of multiple Ito integrals. Therefore, an optimal minimax convergence rate in mean prediction risk of SFLM is valid by using results in literature when the

underlying process  $X(t)$  is a linear combination of multiple Ito integrals. Using the theory of stochastic analysis, one may construct a reproducing kernel Hilbert space (RKHS) associated with a square integrable stochastic process to facilitate analysis of functional data.

### A reproducing kernel Hilbert space framework for functional classification

♦Peijun Sang<sup>1</sup>, Adam Kashlak<sup>2</sup> and Linglong Kong<sup>2</sup>

<sup>1</sup>University of Waterloo

<sup>2</sup>University of Alberta

psang@uwaterloo.ca

We encounter a bottleneck when we try to borrow the strength of classical classifiers to classify functional data. The major issue is that functional data are intrinsically infinite dimensional, thus classical classifiers cannot be applied directly or have poor performance due to curse of dimensionality. To address this concern, we propose to project functional data onto one specific direction, and then a distance-weighted discrimination DWD classifier is built upon the projection score. The projection direction is identified through minimizing an empirical risk function that contains the particular loss function in a DWD classifier, over a reproducing kernel Hilbert space. Hence our proposed classifier can avoid overfitting and enjoy appealing properties of DWD classifiers. This framework is further extended to accommodate functional data classification problems where scalar covariates are involved. In contrast to previous work, we establish a non-asymptotic estimation error bound on the relative misclassification rate. In finite sample case, we demonstrate that the proposed classifiers compare favorably with some commonly used functional classifiers in terms of prediction accuracy through simulation studies and a real-world application.

## Session 17: Recent advances in multivariate and high-dimensional statistics

### Compound Sequential Change Point Detection in Multiple Data Streams

Yunxiao Chen<sup>1</sup> and ♦Xiaoou Li<sup>2</sup>

<sup>1</sup>London School of Economics and Political Science

<sup>2</sup>University of Minnesota

lix1766@umn.edu

We consider sequential change point detection in multiple data streams, where each stream has its own change point. Once a change point is detected for a data stream, this stream is deactivated permanently. The goal is to maximize the normal operation of the pre-change streams, while controlling the proportion of post-change streams among the active streams at all time points. This problem has wide applications in science, social science, and engineering. Taking a Bayesian formulation, we develop a compound sequential decision theory framework for this problem. Under this framework, an oracle procedure is proposed that is optimal among all sequential procedures which control the expected proportion of post-change streams at each time point. We also investigate the asymptotic behavior of the proposed method when the number of data streams grows large. Several non-standard technical tools involving partially ordered spaces and monotone coupling of stochastic processes are developed for proving the optimality result. Numerical examples are provided to illustrate the use and performance of the proposed method.

### Subtask Analysis of Process Data Through a Predictive Model

♦Xueying Tang<sup>1</sup>, Jingchen Liu<sup>2</sup> and Zhiliang Ying<sup>2</sup>

<sup>1</sup>University of Arizona

<sup>2</sup>Columbia University  
xytang@math.arizona.edu

Response process data collected from human-computer interactive items contain rich information about respondents' behavioral patterns and cognitive processes. Their irregular formats as well as their large sizes make standard statistical tools difficult to apply. This paper develops a computationally efficient method for exploratory analysis of such process data. The new approach segments a lengthy individual process into a sequence of short subprocesses to achieve complexity reduction, easy clustering and meaningful interpretation. Each subprocess is considered a subtask. The segmentation is based on sequential action predictability using a parsimonious predictive model combined with the Shannon entropy. Simulation studies are conducted to assess the performance of the new methods. We use the process data from PIAAC 2012 to demonstrate how exploratory analysis of process data can be done with the new approach.

### Spectral clustering via adaptive layer aggregation for multi-layer networks

Sihan Huang<sup>1</sup>, ♦Haolei Weng<sup>2</sup> and Yang Feng<sup>3</sup>

<sup>1</sup>Columbia University

<sup>2</sup>Michigan State University

<sup>3</sup>School of Global Public Health, New York University  
wenghaol@msu.edu

One of the fundamental problems in network analysis is detecting community structure in multi-layer networks of which each layer represents one type of edge information among the nodes. We propose integrative spectral clustering approaches based on effective convex layer aggregations. Our aggregation methods are strongly motivated by a delicate asymptotic analysis of the spectral embedding of weighted adjacency matrices and the downstream k-means clustering, in a challenging regime where community detection consistency is impossible. In fact, the methods are shown to estimate the optimal convex aggregation which minimizes the mis-clustering error under some specialized multi-layer network models. Our analysis further suggests that clustering using Gaussian mixture models is generally superior to the commonly used k-means in spectral clustering. Extensive numerical studies demonstrate that our adaptive aggregation techniques together with Gaussian mixture model clustering make the new spectral clustering remarkably competitive compared to several popularly used methods.

### Detection of Two-Way Outliers in Multivariate Data and Application to Cheating Detection in Educational Tests

♦Yunxiao Chen, Yan Lu and Irini Moustaki

London School of Economics  
y.chen186@lse.ac.uk

This paper concerns the issue of cheating in educational tests due to item leakage. Specifically, we consider the administration of a test, in which some test takers have access to a subset of test items in advance and thus gain advantage. In practice, often both cheating test takers and compromised items are unknown that need to be detected to ensure test fairness. We tackle the simultaneous detection of cheaters and compromised items based on data from a single test administration that consist of item-level binary scores and possibly also item-level response time information. This problem is formulated into an outlier detection problem under a latent variable modeling framework, treating both cheaters and leaked items as outliers. A latent variable model is proposed that adds a latent class model component upon a factor model component, where the factor model component captures normal item response behavior and the latent class model component captures the two-way outliers (i.e.,

cheaters and leaked items). We further propose a statistical decision framework, under which compound decision rules are developed for controlling local false discovery/nondiscovery rates. Statistical inference is carried out under a Bayesian framework, for which a Markov chain Monte Carlo algorithm is developed. The proposed method is applied to data from a computer-based nonadaptive licensure assessment.

## Session 18: Big Data Analysis: New Directions and Innovation

### Sharp Inference on Selected Subgroups in Observational Studies with High Dimensional Covariates

♦Jingshen Wang<sup>1</sup> and Xinzhou Guo<sup>2</sup>

<sup>1</sup>UC Berkeley

<sup>2</sup>Harvard University  
jingshenwang@berkeley.edu

In the study of high-dimensional observational data (e.g., program evaluation, electronic health records data, health care claim database, and educational research), making inference on the maximum treatment effect is particularly helpful for answering some important questions in subgroup analysis, where selecting and making inference on treatment effect for the selected subgroup plays an essential role. Commonly adopted statistical inference applied to the selected subgroup typically assumes the subgroup is chosen independent of data. Such an inferential procedure can lead to an overly optimistic evaluation of the selected subgroup. In the present paper, we propose a resampling framework to simultaneously address the issue of selection bias and the regularization bias induced by the high-dimensional covariates. Our procedure is computationally efficient and provides an asymptotically sharp confidence interval for the maximum treatment effect as well as the selected treatment effect. We demonstrate the merit of our proposal by intensive simulation studies and by analyzing UK Biobank data.

### Sharp Optimality for High Dimensional Covariance Testing

Yumou Qiu

Iowa State University  
yumouqiu@iastate.edu

This paper develops the theoretical limit of testing a high-dimensional covariance being diagonal by deriving the sharp detection boundary as a function of signal proportion and signal strength under alternative hypotheses. The detection boundary gives the exact minimal signal strength that can be detected by some test under the sparse and faint signal regime, which is the most challenging setting for signal detection. We develop an optimal test by multi-level thresholding that is able to achieve the detection boundary. The optimality means the proposed test is powerful as long as the signal strength is above the detection boundary. We establish the asymptotic distribution of the thresholding statistic under non-Gaussian data. A novel  $U$ -statistic composition is developed in conjunction with the matrix blocking and the coupling techniques to handle the complex dependence among sample covariances. We show that the existing tests are non-optimal and the proposed tests are more powerful than those existing tests. Simulation studies are conducted to demonstrate the utility of the proposed test.

### Statistical Inference for Mean Functions of 3D Functional Objects

♦Yueying Wang<sup>1</sup>, Xinyi Li<sup>2</sup>, Guannan Wang<sup>3</sup>, Li Wang<sup>1</sup>, Brandon Klinedinst<sup>1</sup> and Auriel Willette<sup>1</sup>

<sup>1</sup>Iowa State University

<sup>2</sup>University of North Carolina at Chapel Hill

<sup>3</sup>College of William & Mary  
yueyingw@iastate.edu

Functional data analysis has become a powerful tool for conducting statistical analysis for complex objects, such as curves, images, shapes and manifold-valued data. Among these data objects, images obtained using medical imaging technologies emerging recently have been attracting researchers' attention. Examples are functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), which provide a very detailed characterization of brain activity. In general, 3D complex objects are usually collected within the irregular boundary, whereas the majority of existing statistical methods have been focusing on a regular domain. To address this problem, we model the complex data objects as functional data and propose trivariate spline smoothing based on tetrahedralizations for estimating the mean functions of 3D functional objects. The asymptotic properties of the proposed estimator are systematically investigated where consistency and asymptotic normality are established. We also provide a computationally efficient estimation procedure for covariance function and corresponding eigenvalue and eigenfunctions and derive uniform consistency. Motivated by the need for statistical inference for complex functional objects, we then present a novel approach for constructing simultaneous confidence corridors to quantify the uncertainty of the estimation.

#### **Transformation and Integration of Microenvironment Microarray Data**

*Gregory Hunt*

William & Mary  
ghunt@wm.edu

The immediate physical and bio-chemical surroundings of a cell, the cellular microenvironment, is an important component of any fundamental cell and tissue level processes and is implicated in many diseases and dysfunctions. Thus understanding the interaction of cells with their microenvironment can further both basic research and aid the discovery of therapeutic agents. To study perturbations of cellular microenvironments a novel image-based cell-profiling technology called the microenvironment microarray (MEMA) has been recently employed. We explore the effect of preprocessing transformations for MEMA data on the discovery of biological and technical latent effects. We find that Gaussianizing the data and carefully removing outliers can enhance discovery of important biological effects. In particular, these transformations help reveal a relationship between cell morphological features and the extracellular-matrix protein THBS1 in MCF10A breast tissue.

#### **Session 19: Recent Trends of Innovative Methodologies and Applications in Rare Disease Clinical Trials**

##### **A simulation study to evaluate slope model with mixed-model repeated measure for rare disease**

♦ *Tianle Hu and Lixi Yu*<sup>1</sup>

<sup>1</sup>Sarepta Therapeutics  
lhu@sarepta.com

In clinical trials with a continuous endpoint, a mixed-model repeated measure approach (MMRM) often serves as the primary analysis model; However, in rare diseases, the statistical power of such a model may be limited by the small sample size. Therefore, a slope model that leverages the longitudinal data to the fullest extent may be more appealing. Data were simulated for treatment/placebo

from multivariate normal distributions that mimic a type of rare disease. Type I error, power, and bias were evaluated for different scenarios and model specifications.

##### **Adaptive Endpoints Selection with Application in Rare Disease**

♦ *Heng Xu*<sup>1</sup>, *Yi Liu*<sup>1</sup>, *Robert A. Beckman*<sup>2</sup>

<sup>1</sup>nektar therapeutics

<sup>2</sup>Georgetown University Medical Center  
hengxu183@gmail.com

In rare diseases, there are many unanswered questions that are critical to clinical development, such as how to choose primary endpoints that translate into meaningful improvement of health outcomes for patients but at the same time maximize the probability of success. A natural history study is often recommended by regulatory agencies to better understand the natural progression of disease before any pivotal clinical trials are conducted in the specific disease setting. Following this traditional approach has dampened enthusiasm for many drug developers because it entails much higher development cost and longer timeline. We propose to use an innovative design that allows adaptation on primary endpoint(s) so that the learning stage of the disease can be done using information cohort from the pivotal trial itself and no separate natural history study is needed. In disease settings where multiple primary endpoints can be included, we use informational cohort to optimize the alpha allocation among all these endpoints. Otherwise, the informational cohort will be used to select one primary endpoint among the candidates. The decision is primarily based on conditional power and combination test is used under the closed testing principle to ensure no Type I error rate inflation due to the adaptation. Examples and comparisons to the traditional approaches will be presented.

##### **Snapshot Matching: A Method for Borrowing from Longitudinal Historical Control Data**

♦ *Yiyue Lou and Glen Laird*

Vertex Pharmaceuticals  
yiyue.lou@vrtx.com

Randomized clinical trials (RCTs) are considered to be the “gold standard” for evaluating the safety and efficacy of medical treatments. However, for the rare diseases, these trials are not always feasible because of their size, duration, cost, patient preference, or ethical concerns. In this case, there is precedent for borrowing historical control data (e.g., electronic health records, medical claims or patient registry). Propensity score matching is one of the most popular methods for “borrowing”; additional patients from external controls. It matches the treated and control units based on a set of measured covariates to ensure that they are similar in terms of the observable pretreatment characteristics relevant to the disease. When longitudinal historical control data are available, one challenge is to define the index date, i.e. the “baseline”, to establish pretreatment characteristics for the controls. Typically, the first available record in the database is used. However, as the disease progresses, data in later years may be very different from data in earlier years for the same patient regarding key disease progression factors. This may lead to limited numbers of “matchable” historical controls and potentially biased results after matching. It is especially important in rare diseases settings where we would prefer to include all treated patients in the matching and there might be limited historical controls available. We propose a propensity score based matching method called “snapshot matching,” which takes advantage of the longitudinal nature of the historical data, and determines “baseline” for each control unit by the matching algorithm based on its similarity to the treated unit. Key

to this approach is that little background treatment change has occurred over the timeframe of the data, a paradigm not uncommon in rare disease settings. Simulation studies demonstrate that compared with conventional the propensity score matching method, snapshot matching provides better covariate balance after matching, as well as a less biased and less variable treatment effect estimate.

### **BOIN12: Bayesian Optimal Interval Phase I/II Trial Design for Utility-Based Dose Finding in Immunotherapy and Targeted Therapies**

♦ *Ying Yuan, Yahong Zhou<sup>1</sup>, Dianel Li<sup>2</sup>, Fangrong Yan<sup>3</sup> and Ying Yuan<sup>1</sup>*

<sup>1</sup>University of Texas MD Anderson Cancer Center

<sup>2</sup>Bristol-Myers Squibb

<sup>3</sup>China Pharmaceutical University  
rclin@mdanderson.org

For immunotherapy such as checkpoint inhibitors and CAR-T cell therapy, as the efficacy does not necessarily increase with the dose, the maximum tolerated dose (MTD) may not be the optimal dose for treating patients. For these novel therapies, the objective of dose-finding trials is to identify the optimal biological dose (OBD) that optimizes patients' risk-benefit tradeoff. We propose a simple and flexible Bayesian optimal interval phase I/II (BOIN12) trial design to find the OBD that optimizes the risk-benefit tradeoff. The BOIN12 design makes the decision of dose escalation and de-escalation by simultaneously taking account of efficacy and toxicity, and adaptively allocates patients to the dose that optimizes the toxicity-efficacy tradeoff. Compared to existing phase I/II dose-finding designs, the BOIN12 design is simpler to implement, has higher accuracy to identify the OBD, and allocates more patients to the OBD. One of the most appealing features of the BOIN12 design is that its adaptation rule can be pre-tabulated and included in the protocol. During the trial conduct, clinicians can simply look up the decision table to allocate patients to a dose without complicated computation. User-friendly software is freely available at [www.trialdesign.org](http://www.trialdesign.org) to facilitate the application of the BOIN12 design.

### **Session 20: Data Analysis and Application for High-Throughput Biotechnologies**

#### **A powerful genome-wide association test for complex diseases**

*Linchen He and ♦Yongzhao Shao*

New York University School of Medicine  
yongzhao.shao@nyulangone.org

We introduce a powerful genome-wide association test that effectively account for various latent heterogeneity which commonly exists for complex diseases such as Alzheimer's disease, asthma, and cancers. In the context of the widely used case-control type genome-wide association studies (GWAS) for complex diseases, the commonly used association tests have very low power because they do not effectively take into account the latent heterogeneity among cases and among controls. In contrast, the new test have high power at the genome-wide significance level thus can identify a large number of novel variants underlying the heterogeneous etiology of complex diseases. We further propose pine lines and algorithms to prioritize the large number of identified novel variants via eQTL and functional pathway analyses for further mechanistic and translational research. The effectiveness of the proposed method is demonstrated using numerical power analysis and illustrated using GWAS data sets for Alzheimer's disease research.

#### **High-throughput Computational Biology: It's Not Just About the Numbers**

*Robert Nadon*

McGill University  
robert.nadon@mcgill.ca

Computational scientists have embraced the Open Science movement in biomedicine and were among the first to emphasize reproducibility. Naturally, computational scientists have focused on the reproducibility of the numbers produced by their algorithms. If computational science is to be a full participant in the Open Science transformation of 21st-century science, however, the field must go beyond these narrow discipline constraints. When developing algorithms, we must also ask if the output is biologically meaningful. Reproducing algorithm output means little if the numbers mean little. I will present a case study that illustrates how statistical knowledge can help define biological meaningfulness. This case study is also a cautionary tale of the perils of relying solely on life scientists for defining meaningfulness. This talk may particularly interest early-career scientists who are navigating the challenges posed by the current reproducibility crisis in biomedicine.

#### **Untangle Clonal Evolution to Guide Precision Neuro-Oncology Via Data-Intensive Science**

*Jiguang Wang*

Hong Kong University of Science and Technology  
jgwang@ust.hk

Recent advances in next-generation sequencing and data science are revolutionizing numerous areas in life science and medicine. My research is focused on discovering and investigating functional genomic alterations in complex human diseases and relevant biological models by developing and applying computational methods based on statistics and machine learning, aiming to bridge the gaps among data, bench, and bedside. In this talk, I will introduce the latest research on brain cancer evolution and precision medicine in my laboratory. Large-scale genome sequencing projects have uncovered the mutational landscapes of many cancers, but how cancer cells evolve with and without therapy is still unclear. Taking diffuse glioma, the most common and aggressive adult brain cancer, as an example, we aim to address the long-standing questions of how cancer cells respond to therapy and how the founding somatic alterations drive trajectories of cancer evolution.

#### **Issues of z-factor and an approach to avoid them for quality control in high-throughput screening studies**

♦ *Xiaohua Zhang<sup>1</sup>, Dandan Wang<sup>1</sup>, Shixue Sun<sup>1</sup> and Heping Zhang<sup>2</sup>*

<sup>1</sup>University of Macau

<sup>2</sup>Yale University  
douglaszhang@um.edu.mo

High throughput screening (HTS) is a vital automation technology in biomedical research in both industry and academia. The well-known z-factor has been widely used as a gatekeeper to assure assay quality in an HTS study. However, many researchers and users may not have realized that z-factor has major issues. In this presentation, the following four major issues are explored and demonstrated so that researchers may use the z-factor appropriately. First, the z-factor violates the Pythagorean Theorem of Statistics. Second, there is no adjustment of sampling error in the application of the z-factor in HTS studies. Third, the expectation of the sample-based z-factor does not exist. Fourth, the thresholds in the z-factor based criterion lack a theoretical basis. Here, an approach with new criteria to avoid these issues were constructed so that researchers can choose a statistically grounded criterion for quality control (QC) in the HTS

studies. We further implemented this approach in an R package and demonstrated its utility in an HTS study.

## Session 21: Statistical Methods for Sports Data Analytics

### A Bayesian Marked Spatial Point Processes Model for Basketball Shot Chart

♦ *Jieying Jiao, Jun Yan and Guanyu Hu*

University of Connecticut  
jieying.jiao@uconn.edu

The success rate of a basketball shot may be higher at locations where a player makes more shots. In a marked spatial point process, this means the marks are dependent on the intensity. We develop a Bayesian joint model of the mark and the intensity of marked spatial point process, where the intensity is incorporated in the mark's model as a covariate. Further, we allow variable selection through the spike-slab prior. Inferences are developed with a Markov chain Monte Carlo algorithm to sample from the posterior distribution. Two Bayesian model comparison criteria, the modified Deviance Information Criterion and the modified Logarithm of the Pseudo-Marginal Likelihood, are developed to assess the fitness of different models focusing on the mark. The empirical performances of the proposed methods are examined in extensive simulation studies. We apply the proposed methods to the shot charts of four players (Curry, Harden, Durant, and James) in the NBA's 2017–2018 regular season to analyze the shot intensity in the field and the field goal percentage. The results suggest that the field goal percentages of Harden, Durant and James are significantly positively dependent on their shot intensities, while Curry's intensity and field goal percentage are not directly related.

### Grouped Spatial Point Process Model: an Application for Basketball Shot Chart

♦ *Hou-Cheng Yang<sup>1</sup>, Yishu Xue<sup>2</sup> and Guanyu Hu<sup>3</sup>*

<sup>1</sup>Florida State University

<sup>2</sup>Travelers Insurance

<sup>3</sup>University of Connecticut  
hy15e@my.fsu.edu

In this paper, we develop a group linked Log Gaussian Cox process (LGCP) model for analyzing shot pattern of professional basketball players in NBA. We propose a hierarchical Bayesian model for clustering LGCPs based on mixture of finite mixtures (MFM) model. An efficient Markov Chain Monte Carlo (MCMC) algorithm is developed for our proposed model. The empirical performances of the proposed methods are examined in simulation studies, and its usage is further illustrated in analyzing shot charts of several plays in the NBA's 2017–2018 regular season.

### Modeling Quarterback Decision Making in the National Football League

♦ *Matthew Reyers and Tim Swartz*

Simon Fraser University  
mreyers@sfu.ca

This project evaluates quarterback performance in the National Football League. With the availability of player tracking data, there exists the capability to assess various options that are available to quarterbacks and the expected points resulting from each option. The quarterback's execution is then measured against the optimal available option. Since decision making does not rely on the quality of teammates, a quarterback metric is introduced that provides a novel perspective on an understudied aspect of quarterback assessment.

### A Bayesian decision-theoretic approach to uncertain ranks and orderings: Comparing players and lineups

♦ *Andres Barrientos<sup>1</sup>, Deborsee Shen<sup>2</sup>, Garritt Page<sup>3</sup> and David Dunson<sup>4</sup>*

<sup>1</sup>Florida State University

<sup>2</sup>Duke University

<sup>3</sup>Brigham Young University

<sup>4</sup>Duke University  
anfebar@gmail.com

It is common to be interested in rankings or order relationships among entities. In complex settings where one does not directly measure a univariate statistic upon which to base ranks, such inferences typically rely on statistical models having entity-specific parameters. These can be treated as random effects in hierarchical models characterizing variation among the entities. We are particularly motivated by the problem of ranking basketball players in terms of their contribution to team performance. Using data from the United States National Basketball Association (NBA), we find that many players have similar latent ability levels, making any single estimated ranking highly misleading. The current literature fails to provide summaries of order relationships that adequately account for such uncertainty. Motivated by this, we propose a strategy for characterizing uncertainty in inferences on order relationships among players and lineups. Our approach adapts to scenarios in which uncertainty in ordering is high by producing more conservative results that improve interpretability. This is achieved through a reward function within a decision-theoretic framework. We apply our approach to data from the 2009-10 NBA season.

## Session 23: Innovative statistical methods for complex survival data and the applications

### Cancer immunotherapy trial design with long-term survivors

♦ *Jianrong Wu and Xue Ding*

University of Kentucky  
jianrong.wu@uky.edu

Cancer immunotherapy often reflects the improvement in both short-term risk reduction and long-term survival. In this scenario, a mixture cure model can be used for the trial design. However, the hazard functions based on the mixture cure model between two groups will ultimately crossover. Thus, the conventional assumption of proportional hazards may be violated and study design using standard log-rank test (LRT) could lose power if the main interest is to detect the improvement of long-term survival. In this paper, we propose a change sign weighted LRT for the trial design. We derived a sample size formula for the weighted LRT, which can be used for designing cancer immunotherapy trials to detect both short-term risk reduction and long-term survival. Simulation studies are conducted to compare the efficiency between the standard LRT and the change sign weighted LRT.

### Smooth Density Estimation Based on Interval-censored Data with Auxiliary Information

♦ *Qiang Zhao and Martin Schmidt*

Texas State University  
qiang.zhao@txstate.edu

In sexually transmitted disease (STD) studies, the infection time is often interval-censored. Survival probabilities or the probability density function of the infection time can be estimated using Turnbull's self-consistency algorithm or a local likelihood approach with kernel smoothing (Braun and Stafford, 2005). Recently, participants

in some of these studies were asked to keep diaries of their behavioral information. To use this auxiliary information to improve the estimation, a resampling approach was proposed by Harezlak and Tu (2005). In this research, we proposed a kernel density estimation method that incorporates the diary information. Simulation study was conducted to evaluate the proposed method, and comparisons to existing methods were made in terms of bias and mean integrated squared error. For illustration, the proposed method was applied to an STD data for illustration.

#### Group sequential design for historical control trials using error spending functions

Jianrong Wu<sup>1</sup> and Yimei Li<sup>2</sup>

<sup>1</sup>University of Kentucky

<sup>2</sup>St. Jude children's research hospital

yimei.li@stjude.org

Group sequential designs using Lan-DeMets error spending functions are proposed for historical control trials with time-to-event endpoints. Both O'Brien-Fleming and Gamma family types of sequential decision boundaries are derived based on sequential log-rank tests, which follow a Brownian motion in a transformed information time. Simulation results show that the proposed group sequential designs using historical controls preserve the overall type I error and power.

#### On testing the sub-distribution functions under competing risks.

Zhigang Zhang

Memorial Sloan-Kettering Cancer Center

zhangz@mskcc.org

In medical studies, time-to-event endpoints may be subject to competing risks. These competing risks censor each other in a dependent way probabilistically. Usually we are interested in comparing the same cause-specific cumulative incidence or hazard functions across various cohorts but there are situations when we need to compare the cause-specific cumulative incidence or hazard functions across different competing risks. For example, patients treated with radiotherapy may later experience recurrence at different locations and often it is of interest to study such patterns. In this talk I will present an inference procedure for tackling this problem.

### Session 24: Methods and applications in large and complex data

#### High-Dimensional Rank-Based Inference

Xiaoli Kong<sup>1</sup> and Solomon Harrar<sup>2</sup>

<sup>1</sup>Loyola University Chicago

<sup>2</sup>University of Kentucky

xkong1@luc.edu

In this talk, a fully nonparametric (rank-based) method is introduced for comparing multiple groups. To develop the theory, we proved a novel result for studying the asymptotic behavior of quadratic forms in ranks. The simulation study showed that the developed rank-based method performs comparably well with mean-based methods. It has significantly superior power for heavy-tailed distribution with the possibility of outliers. The rank method was applied to EEG data for examining the association between alcohol use and change in brain function.

#### High dimensional change point detection using generalized distance metrics

Shubhadeep Chakraborty and Xianyang Zhang

Texas A&M University

zhangxiany@stat.tamu.edu

Change-point detection has been a classical problem in statistics, finding applications in a wide variety of fields. We consider the problem of nonparametric testing for change-points and estimating the change-point locations for high-dimensional data. Our approach rests upon nonparametric tests for the homogeneity of two high-dimensional distributions. We construct a test statistic based on the new class of homogeneity metrics proposed by Chakraborty and Zhang (2019) and study the asymptotic behavior of our proposed test statistic. Consequently, we illustrate the use of wild binary segmentation to estimate multiple change-point locations hierarchically and provide theoretical consistency results. Finally, we compare the performance of our methodology with other competing methods over simulated and real datasets.

#### A Divide and Conquer Algorithm of Bayesian Density Estimation

Ya Su

University of Kentucky

suyaf@vcu.edu

Data sets for statistical analysis become extremely large even with some difficulty of being stored on one single machine. Even when the data can be stored in one machine, the computational cost would still be intimidating. We propose a divide and conquer solution to density estimation using Bayesian mixture modeling including the infinite mixture case. The methodology can be generalized to other application problems where a Bayesian mixture model is adopted. The proposed prior on each machine or subsample modifies the original prior on both mixing probabilities as well as on the rest of parameters in the distributions being mixed. The ultimate estimator is obtained by taking the average of the posterior samples corresponding to the proposed prior on each subset. Despite the tremendous reduction in time thanks to data splitting, the posterior contraction rate of the proposed estimator stays the same (up to a log factor) as that of the original prior when the data is analyzed as a whole. Simulation studies also justify the competency of the proposed method compared to the established WASP estimator in the finite dimension case. In addition, one of our simulations is performed in a shape constrained deconvolution context and reveals promising results. The application to a GWAS data set reveals the advantage over a naive method that uses the original prior.

#### Nonparametric Methods for Complex Multivariate Data: Asymptotics and Small Sample Approximations

Yue Cui<sup>1</sup> and Solomon Harrar<sup>2</sup>

<sup>1</sup>Missouri State University

<sup>2</sup>University of Kentucky

yuecui@missouristate.edu

Quality of Life (QOL) outcomes are important in the management of chronic illnesses. In studies of efficacies of treatments or intervention modalities, QOL scales—multi-dimensional constructs—are routinely used as primary endpoints. The standard data analysis strategy computes composite (average) overall and domain scores, and conducts a mixed-model analysis for evaluating efficacy or monitoring medical conditions as if these scores were in continuous metric scale. However, assumptions of parametric models like continuity and homoscedasticity can be violated in many cases. Furthermore, it makes it even more challenging when there are missing values on some of the variables. In this talk, we will introduce a purely nonparametric approach in the sense that a meaningful and, yet, nonparametric effect size measures are developed. We propose estimator for the effect size and develop the asymptotic properties.



Our methods are shown to be, particularly effective in the presence of some form of clustering and/or missing values. Inferential procedures are derived from the asymptotic theory. The Asthma Randomized Trial of Indoor Wood Smoke data will be used to illustrate the applications of the proposed methods throughout the talk. The data was collected from a three-arm randomized trial which evaluated interventions targeting biomass smoke particulate matter from older model residential wood stoves in homes that have kids with asthma.

## Session 25: Better Evidence Syntheses in Data Science

### Galaxy plot: a new visualization tool of bivariate meta-analysis studies

*Yong Chen*

University of Pennsylvania  
ychen123@mail.med.upenn.edu

Funnel plots have been widely used to detect small study effects in the results of univariate meta-analyses. However, there is no existing visualization tool that is the counterpart of the funnel plot in the multivariate setting. We present a new visualization method, the galaxy plot, which can simultaneously present the effect sizes of bivariate outcomes and their standard errors in a two-dimensional space. The galaxy plot can be an intuitive visualization tool that can aid in interpretation of results of MMA. It preserves all of the information presented by separate funnel plots for each outcome while elucidating more complex features that may only be revealed by examining the joint distribution of the bivariate outcomes.

### Data fusion using summary versus individual data: relative efficiency for random-effects models

*Dungang Liu*

University of Cincinnati  
dungang.liu@uc.edu

Data fusion is a process of combining information from diverse sources so that a more reliable and efficient conclusion can be reached. It can be conducted by either integrating study-level summary statistics or drawing inference from an overarching model for individual participant data (IPD) if available. The latter is often viewed as the “gold standard”. For random-effects models, however, it remains not fully understood whether the use of IPD indeed gains efficiency over summary statistics. In this paper, we examine the relative efficiency of the two methods under a general likelihood inference setting. We show theoretically and numerically that summary-statistics-based analysis is at most as efficient as IPD analysis, provided that the random effects follow the Gaussian distribution and maximum likelihood estimation is used to obtain summary statistics. More specifically, (i) the two methods are equivalent in an asymptotic sense; and (ii) summary-statistics-based inference can incur an appreciable loss of efficiency if the sample sizes are not sufficiently large. Our results are established under the assumption that the between-study heterogeneity parameter remains constant regardless of the sample sizes, which is different from a previous study. Our findings are confirmed by the analyses of simulated data sets and a real world study of alcohol interventions. This is joint work with Ding-Geng Chen, Xiaoyi Min, and Heping Zhang.

### Estimating the Reference Range from a Meta-analysis

♦ *Lianne Siegel<sup>1</sup>, M. Hassan Murad<sup>2</sup> and Haitao Chu<sup>1</sup>*

<sup>1</sup>University of Minnesota

<sup>2</sup>Mayo Clinic  
siege245@umn.edu

Often clinicians are interested in determining whether a subject’s measurement falls within a normal range, defined as a range of values of a continuous outcome which contains some proportion (eg, 95%) of measurements from a healthy population. Several studies in the biomedical field have estimated reference ranges based on a meta-analysis of multiple studies with healthy individuals. However, the literature currently gives no guidance about how to estimate the reference range of a new subject in such settings. Instead, meta-analyses of such normative range studies typically report the pooled mean as a reference value, which does not incorporate natural variation across healthy individuals in different studies. We present three approaches to calculating the normal reference range of a subject from a meta-analysis of normally or lognormally distributed outcomes: a frequentist random effects model, a Bayesian random effects model, and an empirical approach. We present the results of a simulation study demonstrating that the methods perform well under a variety of scenarios, though users should be cautious when the number of studies is small and between-study heterogeneity is large. Finally, we apply these methods to two examples: pediatric time spent awake after sleep onset and frontal subjective postural vertical measurements.

### Predictive treatment ranking in Bayesian network meta-analysis

*Lifeng Lin*

Florida State University  
llin4@fsu.edu

Network meta-analysis (NMA) is an important tool to provide high-quality evidence about available treatments’ benefits and harms for comparative effectiveness research. Compared with conventional meta-analyses that synthesize related studies for pairs of treatments separately, an NMA uses both direct and indirect evidence to simultaneously compare all available treatments for a certain disease. It is of primary interest for clinicians to rank these treatments and select the optimal ones for patients. Various methods have been proposed to evaluate treatment ranking; among them, the mean rank and the so-called surface under the cumulative ranking curve (SUCRA) are widely used in current practice of NMAs. However, these measures only summarize treatment ranks among the studies collected in the NMA; due to heterogeneity between studies, they cannot predict treatment ranks in a future study and thus may not be directly applied to healthcare for new patients. In this talk, we propose innovative measures to predict treatment ranks by accounting for the heterogeneity between the existing studies in an NMA and a new study. They are the counterparts of the mean rank and the SUCRA under the new study setting. We use two illustrative examples and simulations to evaluate the performance of the proposed measures.

## Session 26: Deciphering Multi-omics Data: Statistical Models and Computational Approaches for Biology and Health

### Cluster Ensemble and Batch Effect Correction Methods for Single Cell RNA-sequencing Data

*Yun Li*

University of North Carolina  
yun.li@med.unc.edu

Single-cell RNA sequencing (scRNA-seq) allows researchers to examine the transcriptome at the single-cell resolution and has been increasingly employed as technologies continue to advance. Due to

technical and biological reasons unique to scRNA-seq data, clustering and batch effect correction are almost indispensable to ensure valid and powerful data analysis. Multiple methods have been proposed for these two important tasks. For clustering, we have found that different methods, including state-of-the-art methods such as Seurat, SC3, CIDR, SIMLR, t-SNE + k-means, yield varying results in terms of both the number of clusters and actual cluster assignments. We have developed ensemble methods, SAFE-clustering and SAME-clustering, that leverages hyper-graph partitioning algorithms and a mixture model-based approach respectively to produce more robust and accurate ensemble solution on top of clustering results from individual methods. For batch effect correction, we have developed methods based on supervised mutual nearest neighbor detection to harness the power of known cell type labels for certain single cells. We benchmarked all methods in various scRNA-seq datasets to demonstrate their utilities.

#### **Fine association testing for whole genome sequencing data with knockoffs**

Zihuai He

Stanford University  
zihuai@stanford.edu

The analysis of whole-genome sequencing studies is challenging due to a large number of noncoding rare variants, our limited understanding of their functional effects, and the lack of natural units for testing. Existing methods based on FWER control and marginal association tests often suffer from insufficient power owing to the expansion of the multiple testing problem, and increased risk of false positives due to the presence of linkage disequilibrium. We propose a scan statistic framework and a sequential knockoff generator for whole-genome sequencing data analysis. The proposed method is able to simultaneously detect the existence and localizes association signals at genome-wide scale with guaranteed FDR control, and to significantly reduce false discoveries due to moderate linkage disequilibrium.

#### **Causal Inference for Heritable Phenotypic Risk Factors Using Heterogeneous Genetic Instruments**

♦Jingshu Wang<sup>1</sup>, Qingyuan Zhao<sup>2</sup>, Jack Bowden<sup>3</sup>, Gibran Hemani<sup>4</sup>, George D. Smith<sup>5</sup>, Dylan S. Small<sup>6</sup> and Nancy R. Zhang

<sup>1</sup>University of Chicago

<sup>2</sup>University of Cambridge

<sup>3</sup>University of Exeter

<sup>4</sup>University of Bristol

<sup>5</sup>University of Bristol

<sup>6</sup>University of Pennsylvania  
wangjingshususan@gmail.com

Over a decade of genome-wide association studies have led to the finding that significant genetic associations tend to be spread across the genome for complex traits, leading to the recent proposal of an "omnigenic" model where almost all genes contribute to every complex trait. Such an omnigenic phenomenon complicates Mendelian Randomization studies, where natural genetic variations are used as instruments to infer the causal effect of heritable risk factors. We reexamine the assumptions of existing Mendelian Randomization methods and show how they need to be revised to allow for pervasive pleiotropy and heterogeneous effect sizes. We propose a comprehensive framework GRAPPLE (Genome-wide mR Analysis under Pervasive PLEiotropy) to analyze the causal effect of target risk factors with heterogeneous genetic instruments and identify possible pleiotropic patterns from data. By using summary statistics from genome-wide association studies, GRAPPLE can efficiently

use both strong and weak genetic instruments, detect the existence of multiple pleiotropic pathways, adjust for confounding risk factors, and determine the causal direction. With GRAPPLE, we analyze the effect of blood lipids, body mass index, and systolic blood pressure on 25 disease outcomes, gaining new information on their causal relationships and the potential pleiotropic pathways.

#### **HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity**

Shuang Jiang<sup>1</sup>, Guanghua Xiao<sup>2</sup>, Andrew Koh<sup>2</sup>, Yingfei Chen<sup>3</sup>, Bo Yao<sup>2</sup>, Qiwei Li<sup>4</sup> and ♦Xiaowei Zhan

<sup>1</sup>Southern Methodist University

<sup>2</sup>University of Texas Southwestern Medical Center

<sup>3</sup>University of Texas

<sup>4</sup>The University of Texas at Dallas  
xiaowei.zhan@utsouthwestern.edu

The human microbiome is a collection of microorganisms. They form complex communities and collectively affect host health. Recently, the advances in next-generation sequencing technology enable the high-throughput profiling of the human microbiome. This calls for a statistical model to construct microbial networks from the microbiome sequencing count data. As microbiome count data are high-dimensional and suffer from uneven sampling depth, overdispersion, and zero-inflation, these characteristics can bias the network estimation and require specialized analytical tools. Here we propose a general framework, HARMONIES, a Hybrid Approach for Microbiome Network Inferences via Exploiting Sparsity, to infer a sparse microbiome network. HARMONIES first utilizes a zero-inflated negative binomial (ZINB) distribution to model the skewness and excess zeros in the microbiome data, as well as incorporates a stochastic process prior for sample-wise normalization. This approach infers a sparse and stable network by imposing non-trivial regularizations based on the Gaussian graphical model. In comprehensive simulation studies, HARMONIES outperformed four other commonly used methods. When using published microbiome data from a colorectal cancer study, it discovered a novel community with disease-enriched bacteria. In summary, HARMONIES is a novel and useful statistical framework for microbiome network inference, and it is available at <https://github.com/shuangj00/HARMONIES>.

#### **Session 27: Advanced Adaptive Enrichment Designs in Confirmative Clinical Trials**

##### **A Case Study of Adaptive Population Enrichment Design in a Phase 3 Oncology Trial**

Bo Jin

Boston Biomedical, Inc  
bjin@bostonbiomedical.com

Clinical trials are traditionally conducted in a general population while an ad-hoc subset analysis is performed at the final analysis to determine efficacy in subsets of patients. Such an approach has been ethically challenging and is inefficient to assess drug efficacy when the drug is found to be effective only in a subgroup of patients, e.g., who are associated with a newly discovered biomarker. In recent years, the adaptive population enrichment framework has successfully adopted data-driven decision rules to define adaptive population selection. In this presentation, we will present a case study of Phase 3 oncology trial in which both general population and a biomarker-defined sub-population are studied, and an interim analysis is employed to evaluate the treatment effect

within each of the two population at the interim look and identify the most promising populations. The final analysis is conducted within the selected populations using the data collected before and after the interim analysis. Other adaptive design rules are also be considered including futility analysis and sample size re-estimation. The combination principle is employed to guarantee overall Type I error.

#### **Complex multiplicity problems in adaptive designs with population selection.**

♦ *George Kordzakhia and Alex Dmitrienko*

FDA

george.kordzakhia@fda.hhs.gov

Adaptive clinical trials are often designed to pursue complex sets of objectives with data-driven decisions at interim looks. This talk presents a general framework for setting up multiple testing procedures for complex multiplicity problems arising in clinical trials with adaptive designs that support population selection at an interim analysis. The procedures are defined using the closed testing principle and, to account for multiple data-driven decisions, the inverse normal combination function approach is applied. The resulting multiplicity adjustment framework is flexible and can be applied to a broad class of adaptive designs.

#### **Practical considerations for adaptive enrichment design implementation**

♦ *Jianchang Lin, Sheela Kolluri, Veronica Bunn and Rachael Liu*

Takeda Pharmaceuticals

jianchang.lin@takeda.com

In the current precision medicine era, there has been an explosion in the knowledge of the molecular profile of diseases. This provides vast opportunities to conduct clinical trials in the setting of biomarkers and targeted therapies, where the traditional paradigm of treating very large number of unselected patients is increasingly less efficient, lack of cost effectiveness and ethically challenging. Efficient clinical trial designs to establish the treatment efficacy in pre-specified subgroups and the overall population are critical in these setting. We proposed a clinical trial design combines Bayesian decision tools for subgroup selection and adaptive design with sample size re-estimation to optimize risk mitigation. This design provides the option to rescue a trial at a later interim analysis with more mature data and improves the probability of success when there are heterogeneous subgroup effects. The approach could allow gains in the timeline and the number of patients with seamless development strategy, since that it does not need to wait for the completion of independent phase II results before the initiation of the phase III trials. We also conduct simulations to illustrate the operating characteristics of different methodological decision tools and designs, along with practical considerations for adaptive enrichment design implementation in real case studies.

#### **Discussant**

*Jared Christensen*

NA

jared.christensen@pfizer.com

Discussant

### **Session 28: New Challenges and Opportunities in Early-Phase Oncology Trials**

#### **Practical Considerations in Implementing Modern Dose-Escalation Methods from an Industry's Perspective**

♦ *Yuanyuan Bian, Aimee Wang and Wei Zhang*

Eli Lilly and Company

bian.yuanyuan@lilly.com

Dose-escalation (DE) in early-phase oncology trials is ever-increasingly challenging in recent years: on one hand, novel oncologic agents with different mechanisms of action could result in drastically different toxicity profiles such as no dose-limiting toxicity (DLT) or delayed toxicity; on the other hand, industry sponsors, facing fierce competitions, become even more pressed to expedite the DE process to get an optimal dose for the next stage development. Innovative statistical methods have been proposed to address such emerging challenges: for late-onset toxicity and potential shortening of trial duration (e.g., enroll patients with pending DLT outcomes), new time-to-event (TITE)-based methods are developed based on model-assisted methods, such as TITE-BOIN/keyboard, R-TPI, and PoD-TPI; models incorporating toxicity burden with not only conventional DLTs but also lower-grade toxicities are being built (e.g., TITE-ToBI); there is also an increasing trend of sponsors utilizing single-patient acceleration or titration designs to expedite the process. These approaches address key questions from various aspects to improve the identification of maximum tolerated dose (MTD) and may effectively reduce trial duration. In this talk, we will discuss some practical considerations and issues that are likely to arise when implementing these new methods and our view regarding the decision of recommended phase 2 dose (RP2D). The selected MTD alone does not necessarily equate RP2D, as the RP2D decision may also be heavily influenced by the totality of data and sometimes empirical medical experience. Subtle but critical aspects not currently considered among the novel methods for RP2D determination deserve to be addressed in a more quantitative and rigorous manner. For instance, the inclusion of "backfill" patients, characteristics of adverse events, impact of poorly pre-defined dose range/increments on the operating characteristics of MTD determination, higher rate of dose adjustment beyond DLT-evaluation period, etc., will be discussed in this talk with possible directions for solutions.

#### **Titration of T-cell Engager (TiTE): A new method for Phase 1 dose-finding design for systematic intra-subject dose escalation with application to T-cell Engagers**

♦ *Chenjia Xu<sup>1</sup>, Bin Zhuo<sup>2</sup> and Erik Rasmussen<sup>2</sup>*

<sup>1</sup>Indiana University

<sup>2</sup>Amgen Inc.

cx4@iu.edu

T-cell engagers are a class of oncology drugs which engage T-cells to initiate immune response against malignant cells. T-cell engagers have features that are unlike prior classes of oncology drugs (e.g., chemotherapies or targeted therapies), because 1) starting dose level often must be conservative due to immune-related side effects such as cytokine release syndrome; 2) dose level can be titrated to higher as a result of subject's immune system adaptation after first exposure to lower dose; and 3) dose limiting toxicities are rarely observed. It is generally believed that for T-cell engagers the dose intensity of the starting dose level and the peak dose intensity both correlate with improved efficacy. Existing dose finding methodologies, such as Bayesian logistic regression model (BLRM), are not designed to efficiently identify both the initial starting dose and peak dose intensity in a single trial. In this study, we propose a framework that can 1) estimate the maximum tolerated initial dose level (MTD1); and 2) incorporate systematic intra-subject dose-escalation to estimate the maximum tolerated dose level subsequent to the initial dose level (MTD2) with a survival analysis approach. We compare our framework to similar methodologies and evaluate

their key operating characteristics.

### **uTPI: A Utility-Based Toxicity Probability Interval Design for Dose Finding in Phase I/II Trials**

*Ruitao Lin*

The University of Texas MD Anderson Cancer Center  
rlin@mdanderson.org

Molecularly targeted agents and immunotherapy have revolutionized modern cancer treatment. Unlike chemotherapy, the maximum tolerated dose of the targeted therapies may not pose significant clinical benefit over the lower doses. By simultaneously considering both binary toxicity and efficacy endpoints, phase I/II trials can identify a more clinically meaningful dose for subsequent phase II trials than traditional phase I trials in terms of risk-benefit tradeoff. Existing phase I/II dose-finding methods are model-based or need to pre-specify many design parameters, which makes them difficult to implement in practice. To strengthen and simplify the current practice of phase I/II trials, we propose a utility-based toxicity probability interval (uTPI) design for finding the optimal biological dose (OBD) where binary toxicity and efficacy endpoints are observed. The uTPI design is model-assisted in nature, directly modeling the utility outcomes observed at the current dose level based on a quasi binomial likelihood. Toxicity probability intervals are used to screen out overly toxic dose levels, and then the dose escalation/de-escalation decisions are made adaptively by comparing the posterior utility distributions of the adjacent levels of the current dose. The uTPI design is flexible in accommodating various utility functions while only needs minimum design parameters. A prominent feature of the uTPI design is that it has a simple decision structure such that a concise dose-assignment decision table can be calculated before the start of trial and be used throughout the trial, which greatly simplifies practical implementation of the design. Extensive simulation studies demonstrate that the proposed uTPI design yields desirable as well as robust performance under various scenarios.

### **Session 29: Novel clinical trial designs in the era of precision medicine and immunotherapy**

#### **TITE-BOIN-ET: Time-to-event Bayesian optimal interval design to accelerate dose-finding based on both efficacy and toxicity outcomes**

*Kentaro Takeda*

Astellas  
kentaro.takeda@astellas.com

One of the primary purposes of an oncology dose-finding trial is to identify an optimal dose (OD) that is both tolerable and has an indication of therapeutic benefit for subjects in subsequent clinical trials. In addition, it is quite important to accelerate early-stage trials to shorten the entire period of drug development. However, it is often challenging to make adaptive decisions of dose escalation and de-escalation in a timely manner because of the fast accrual rate, the difference of outcome evaluation periods for efficacy and toxicity and the late-onset outcomes. To solve these issues, we propose the time-to-event Bayesian optimal interval design to accelerate dose-finding based on cumulative and pending data of both efficacy and toxicity. The new design, named “TITE-BOIN-ET” design, is non-parametric and a model-assisted design. Thus, it is robust, much simpler, and easier to implement in actual oncology dose-finding trials compared with the model-based approaches. These characteristics are quite useful from a practical point of view. A simulation study shows that the TITE-BOIN-ET design has advantages com-

pared with the model-based approaches in both the percentage of correct OD selection and the average number of patients allocated to the ODs across a variety of realistic settings. In addition, the TITE-BOIN-ET design significantly shortens the trial duration compared with the designs without sequential enrollment and therefore has the potential to accelerate early-stage dose-finding trials.

#### **Novel Early Phase Clinical Trial Designs for Cancer Therapeutic Vaccines**

*Chenguang Wang*

Johns Hopkins University  
cwang68@jhmi.edu

Cancer vaccines that treat existing pre-cancer or cancer are known as cancer therapeutic vaccines. When testing cancer vaccines in cancer patients, investigators often face the hurdle that typical clinical trial designs for cytotoxic drugs are not fully applicable. Therefore, there is a need for statisticians to bring forth innovative clinical trial designs for evaluating the safety and efficacy of cancer vaccines. In this talk, we will talk about the challenges and possible solutions for designing and conducting clinical trials in the development of cancer vaccines.

#### **Hierarchical Bayesian Clustering Design of Multiple Biomarker Subgroups (HCOMBS)**

♦ *Jun (Vivien) Yin, Daniel Kang<sup>1</sup> and Qian Shi<sup>2</sup>*

<sup>1</sup>University of Iowa

<sup>2</sup>Mayo Clinic  
yin.jun@mayo.edu

The paradigm for cancer clinical trials has shifted towards individualized treatment due to the blooming discoveries of biomarkers and targeted agents. Because of the deficiencies of screening agents or testing histologic tumor types one at a time, basket and umbrella trials are emerging. We proposed Hierarchical Bayesian Clustering Design of Multiple Biomarker Subgroups (HCOMBS) for designing and conducting umbrella trials, to evaluate biomarker-treatment pairing and identify patient subpopulations that are most likely to benefit from novel agents. Compared to parallel design for individual cohorts, the HCOMBS designs have greatly reduced sample size, and hence improve efficiency and decreases the financial costs. We further extended it to allow the parallel cohorts to be dynamic, so that at interim analysis investigators can 1) eliminate or graduate molecular subgroups, and 2) cluster subgroups that exhibit similar response to treatment to improve power. The designs were calibrated with respect to specific error rates. We conduct extensive simulations to assess the performance, and illustrate with a genomically-guided treatment trial in brain metastases (A071701) conducted by the NCI cooperative group Alliance for Clinical Trials in Oncology.

### **Session 30: Statistical inference and practical issues in psychiatry**

#### **Statistical Ethics in Psychiatry**

*Jane Kim*

Stanford University  
janepkim@stanford.edu

As health care systems increasingly utilize algorithms for patient identification, diagnosis, and treatment direction for psychiatric conditions, the consequences of algorithmic weakness, such as bias, yield real and significant costs. Machine learning-driven algorithmic medicine now faces an urgent need to anticipate and address emerging ethical issues. In this talk I will introduce “empirical

ethics” as a necessary component to address this problem and secondly, the potential for statistical applications to benefit this area. Empirical ethics inquiry works from the notion that stakeholder perspectives are necessary for gauging the ethical acceptability of the application of ML in healthcare, and assuring that this work aligns with societal expectations. We will present a few possibilities for statistical methods to strengthen the area of ethics.

### Latent Class Mediator

Haiqun Lin

Rutgers University  
haiqun.lin@yahoo.com

This study demonstrates the utility of latent classes to study the effect of an intervention on an outcome through multiple indicators of mediation. These indicators are regarded as observed intermediate variables that identify an underlying latent class mediator with each class representing a different mediating pathway. We adopt the potential outcome framework to estimate the mediating effect of each latent class. The use of a latent class mediator ensures the decomposition of the total mediating effect into additive effects from individual mediating pathways, a highly desirable feature for multiple mediators. A simultaneous estimation approach through the maximum likelihood is proposed. This method is applied to the analysis of the first six months of data from a two-year clustered randomized clinical trial for young adults in their first episode of schizophrenia, the Recovery After an Initial Schizophrenia Episode Early Treatment Program. The four indicators of mediation are considered: individual resiliency training; family psychoeducation; supported education and employment; and a structural assessment for medication. The improvement in symptoms of schizophrenia at the sixth month was found to be mediated by the latent class mediator derived from these four indicators of service. Simulation studies were conducted to assess the performance of the latent class mediation model and showed that the simultaneous estimation yielded little bias when the entropy of the indicators was high.

### LONGITUDINAL CANONICAL CORRELATION ANALYSIS

◆Seonjoo Lee<sup>1</sup>, Jongwoo Choi<sup>2</sup> and Zhiqian Fang<sup>1</sup>

<sup>1</sup>Columbia University and New York State Psychiatric Institute

<sup>2</sup>New York State Psychiatric Institute  
sl3670@cumc.columbia.edu

This talk considers canonical correlation analysis for the two longitudinal variables that are possibly sampled at different time resolutions with irregular grids. We modeled trajectories of the multivariate variables using random effects and found the most correlated sets of linear combinations in the latent space. The numerical simulation showed that the LCCA could recover underlying correlation patterns between two high-dimensional longitudinal data. The proposed methods were applied to Alzheimer’s Disease Neuroimaging Initiative Data and identified the longitudinal profiles of morphological brain changes and amyloid cumulation.

### Deep Neural Network for Interval-Censored Survival Outcome Using Genetic Data, with an Application to Predict AD Progression

Tao Sun<sup>1</sup> and Ying Ding<sup>2</sup>

<sup>1</sup>Renmin University

<sup>2</sup>University of Pittsburgh, USA  
yingding@pitt.edu

Informative and accurate prediction with individualized risk of progression profiles over time is critical for personalized intervention/treatment and clinical management. The massive genetic data,

such as SNPs from genome-wide association studies (GWAS), together with well-characterized time-to-event phenotypes provide unprecedented opportunities for developing effective survival prediction models. Motivated by developing accurate prediction models for the progression of Alzheimer’s Disease (AD), where the progression time are interval-censored (IC) due to intermittent assessment times, we first develop a semiparametric transformation model to flexibly model the IC data and to successfully identify top SNPs that are associated with AD progression. Then we propose and implement a multilayer deep neural network (DNN) survival model for interval-censored data to effectively extract features and make accurate and interpretable predictions. Finally, using the GWAS data from Alzheimer’s Disease Neuroimaging Initiative (ADNI), we apply our DNN-IC survival model with a large number of top SNPs from GWAS to establish an accurate predictive model for AD progression. Moreover, we obtain a subject-specific importance measure for each predictor from the DNN survival model, which provides valuable insights into the personalized early prevention and clinical management for this disease.

### Session 31: Recent development in dynamic historical data borrowing: methodology and application in clinical trials

#### A Dynamic Frequentist Approach of Historical Data

◆Bingming Yi and Chenghao Chu

Vertex Biopharmaceuticals  
bingming.yi@vrtx.com

In clinical trials in rare diseases or a certain population like pediatric, it is of great interest to incorporate historical data to increase power of evaluating the treatment effect of an experimental drug. However, it remains a challenge to control type-1 error rate or inflate it to a tolerable degree with a desirable power. To address this issue, dynamic historical borrowing approaches borrow historical data more when historical data is similar to current and less otherwise. This talk proposed to use a weighted average of historical and current control data as new control, with the weight being set as an approximation to the optimal weight that minimizes the mean squared errors in the treatment effect estimation. Comparing to selected existing methods, the proposed method was less subjective, and showed superior power and adequate type-1 error control through simulation studies. Keywords: Clinical trial, historical data borrowing, rare disease, study design.

#### The promises and compromises of dynamic borrowing in clinical trials

◆Xiaodong Luo and Hui Quan

Sanofi  
xiaodong.luo@sanofi.com

There are large amounts of clinical data available before the start of a study, particularly on the current study’s control arm. It is therefore appealing to “borrow” this information so that the current study can focus more on the novel treatment while retaining accurate estimates of the current control arm parameters. Fixed borrowing of historical data is simple and easily interpretable however it does not guard against potential inconsistency between the historical data and current data. Dynamic borrowing with the amount of borrowing depending on the level of consistency between the historical and current data has its advantage of being data adaptive. The downside is that such approach often requires extensive simulation and the results may not be easily understandable. In this talk, we will present

some recent thinking/research on dynamic historical data borrowing, in particular, point out the promises and compromises of using such approach in clinical trials.

#### **Adaptive Conditional Borrowing of Historical Data in Rare Disease Development**

◆ *Yingying Liu<sup>1</sup>, Peng Sun<sup>1</sup>, Charlie Cao<sup>1</sup>, Bo Lu<sup>2</sup>, Ming-Hui Chen<sup>3</sup>, John Zhong<sup>4</sup>, Richard Foster, Susie Sinks<sup>1</sup>, Fan Wu<sup>1</sup> and Giulia Gambino<sup>1</sup>*

<sup>1</sup>Biogen

<sup>2</sup>Ohio State University

<sup>3</sup>University of Connecticut

<sup>4</sup>REGENXBIO

yingying.liu@biogen.com

Historical data borrowing has often been used in drug development to increase the probability of success and reduce patient burden, especially for the treatment of serious rare diseases. However, improper use of historical data can cause biased treatment effect estimation and inflated type I error. Thus, for a study that attempts to borrow historical data, the similarity between the historical control and concurrent treatment populations is an important consideration. Previous research on conditional power prior approach showed some appealing characteristics in terms of bounded type I error and unbiased treatment effect estimation. In this presentation, we will present a study design that optimizes the propensity score methods and conditional borrowing approach to ensure the comparability of responses between historical and concurrent control, in addition to matching key baseline covariates and SOC. Different matching methods will be discussed and simulation studies will be carried out to examine the operational characteristics.

### **Session 32: Bayesian Analysis of Complex Survey Data**

#### **Statistical Integration and Inference via Multilevel Regression and Poststratification**

*Yajuan Si*

University of Michigan  
yajuan@umich.edu

We develop a unified framework under multilevel regression and poststratification (MRP) for data integration and inferences. In this article we demonstrate the capability of MRP to handle the methodological and computational issues on big data in the combination of probability and nonprobability-based surveys. The emergence of big data provides unprecedented resources for population-based studies to address policy-related questions. However, such data may not be representative of the target population as convenience or volunteer samples, a form of nonprobability-based selection. Nonprobability samples become popular with the quick collection and low cost, in contrast with the rapidly declining response rate and increasing cost of probability surveys, which leads to a new direction of survey research. Data integration and record linkage become research priorities for most statistical agencies. The lack of theoretical foundations under new data collection methods presents challenges to traditional design-based approaches. As a promising solution with influential applications, MRP stabilizes small area estimation and accounts for the sample selection and response mechanisms into modeling. MRP can predict outcome values for nonsampled units and propagate all sources of uncertainty. We use simulation studies to evaluate the frequentist properties and statistical validity of MRP in comparison with alternative methods.

#### **Fully Bayesian Estimation under Dependent and Informative Cluster Sampling**

*Luis Leon Novelo*

The University of Texas Health Science Center at Houston

Luis.G.LeonNovelo@uth.tmc.edu

Survey data are often collected under multistage sampling designs where units are binned to clusters that are sampled in a first stage. The unit-indexed population variables of interest are typically dependent within cluster. We propose a Fully Bayesian method that constructs an exact likelihood for the observed sample to incorporate unit-level marginal sampling weights for performing unbiased inference for population parameters while simultaneously accounting for the dependence induced by sampling clusters of units to produce correct uncertainty quantification. Our approach parameterizes cluster-indexed random effects in both a marginal model for the response and a conditional model for published, unit-level sampling weights. We compare our method to plug-in Bayesian and frequentist alternatives in a simulation study and demonstrate that our method most closely achieves correct uncertainty quantification for model parameters, including the generating variances for cluster-indexed random effects. We demonstrate our method in one application with NHANES data.

#### **Bayesian bivariate models for identifying common spatial patterns in small area estimation using survey data with weights**

*Cici Bauer*

The University of Texas Health Science Center at Houston

cici.x.bauer@uth.tmc.edu

Bayesian shared component model has been widely used in identifying the common spatial pattern among different outcomes. In many cases, when one or more of the data sources used in such models are from survey, it is important to include the sampling weights to reflect the complex surveys. In this talk, I will describe an extension of the Bayesian shared component model that allows incorporating the sampling weights for binary data collected from survey. We demonstrated the utility of our proposed approach in understanding the common spatial pattern of COVID-19 cases and diabetes prevalence, using the Cameron County Hispanic Cohort data and the reported COVID-19 data. This is joint work with faculty members from UTSPH-Brownsville.

#### **Locally Adaptive Shrinkage in Generalized Linear Models**

◆ *Andrew Womack<sup>1</sup> and Daniel Taylor-Rodriguez<sup>2</sup>*

<sup>1</sup>Indiana University

<sup>2</sup>Portland State University

ajwomack@indiana.edu

Bayesian locally adaptive shrinkage models are a class of models that approximate the discrete selection problem through the use of continuous shrinkage parameters instead of discrete inclusion parameters. In the context of Gaussian models, this is achieved by a continuous interpolation between a point mass at the origin and the Lebesgue measure on Euclidean space. In generalized linear models, “flatness” for regression coefficients is not represented by the Lebesgue measure, but rather by some default or reference prior. We specify locally adaptive shrinkage models that interpolate between a point mass and some default prior and explore their model selection properties.

### Session 33: Multiple phenotypes, Pleiotropy and Mendelian Randomization

#### Identifying pleiotropic loci between type 2 diabetes and prostate cancer

♦*Debashree Ray and Nilanjan Chatterjee*

Johns Hopkins University  
dray@jhu.edu

With the growing number of disease- and trait-associated genetic variants consistently detected and replicated across many genome-wide association studies (GWAS), there is increasing evidence that pleiotropy (the association of multiple traits with same genetic variants/loci) is a very common phenomenon. Cross-phenotype association tests are being used more often to jointly analyze multiple traits from a GWAS. The underlying methods, however, are often designed to test the global null hypothesis that there is no association for a genetic variant across any of the trait, the rejection of which does not imply pleiotropic association. In particular for case-control traits, the typical approaches for identifying genetic overlap between diseases either involve finding common significant findings from genome-wide analysis of each disease separately (which ignores the joint variation of the diseases), or investigating how well polygenic risk score of one explains the variation of the other (which does not implicate novel pleiotropic variants). We propose a novel approach for detecting pleiotropic loci across two traits by considering an underlying null hypothesis that a variant is associated with none or only one of the traits. We perform extensive simulations to study properties such as type I error and power under different scenarios. Application of our approach to publicly available summary data from two of the largest case-control GWAS of Type 2 Diabetes and of Prostate Cancer not only replicated known and candidate shared genes but also implicated a few potentially novel shared genetic regions.

#### Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects

♦*Guanghao Qi and Nilanjan Chatterjee*

Johns Hopkins University  
gqi1@jhmi.edu

Mendelian Randomization (MR) has emerged as a major tool for the investigation of causal relationship among traits, utilizing results from large-scale genome-wide association studies. Bias due to “horizontal pleiotropy”, however, remains a major concern. We propose a novel approach for robust and efficient MR analysis using large number of genetic instruments, based on a novel spike-detection algorithm under a normal-mixture model for underlying effect-size distributions. Simulations show that the new method, MRMix, provides nearly unbiased or/and less biased estimates of causal effects compared to alternative methods and can achieve higher efficiency than comparably robust estimators. Application of MRMix to publicly available datasets leads to notable observations, including identification of causal effects of BMI and age-at-menarche on the risk of breast cancer; no causal effect of HDL and triglycerides on the risk of coronary artery disease; a strong detrimental effect of BMI on the risk of major depressive disorder.

#### Pleiotropy and Mendelian Randomization analysis using GWAS summary statistics

*Xiaofeng Zhu*

Case Western Reserve university  
xxz10@case.edu

The overall association evidence of a genetic variant with multiple traits can be evaluated by cross phenotype association analysis using

summary statistics from genome wide association studies (GWAS). Further dissecting the association pathways from a variant to multiple traits is important to understand the biological causal relationships among complex traits. Here I will introduce a flexible and computationally efficient Iterative Mendelian Randomization and pleiotropy (IMRP) approach to search horizontal pleiotropic variants and estimate causal effect simultaneously. I will use both simulations and real data analysis to illustrate this method and compare with existing methods.

### Session 34: New methods of clinical trial designs and analyses and sample size re-estimation

#### Some Discussions on Sample Size Re-estimation

♦*Victoria Chang, Jianfei Zheng and Yan Ma*

BeiGene  
changvick@gmail.com

One of the critical steps in clinical trial design is to have an adequate sample size for a sufficient powered study. A sample size re-estimation design gives the flexibility to reevaluate the accuracy on the assumption of the parameters required for the sample size calculation prior to the trial. This is also the adaptive design approach accepted by many regulatory agencies. This presentation will give a brief overview on the recent development of sample size re-estimation and the guidelines from regulatory agencies.

#### A Robust Design Approach for Clinical Trials with Potential Nonproportional Hazards: A Straw Man Proposal

*Satrajit Roychoudhury*

Pfizer Inc.  
satrajit.roychoudhury@pfizer.com

Targeting the immune system to cure cancer has emerged as a promising treatment option for patients in recent years. Instead of targeting a tumor directly or destroying it with radiation, Immunotherapy boosts the body’s natural defenses to fight cancer. However, this novel treatment poses new challenges in the study design and statistical analysis of clinical trials. A major challenge is the delayed onset of treatment effects due to the mechanism of immunotherapy which violates the proportional hazard (PH) assumption. The conventional log-rank test may suffer a significant power loss in such scenarios. It is known as the non-proportional hazard (NPH) problem. A suitable design for time to event data with potential NPH needs to be flexible enough to incorporate the uncertainty of NPH type and provide a robust inference. This presentation will focus on an alternative design approach for immune-oncology trials. The proposed design approach is based on a combination of multiple Fleming-Harrington WLR tests and is referred as the Max-Combo test. It chooses the best test adaptively depending on the underlying data. The main objective of the new design is to provide robust power for primary analysis under different NPH scenarios. The talk will provide the general design framework, sample size calculation, and evaluation of operating characteristics. In addition, a comparison of MaxCombo with other available approaches will be provided. Finally, It will reflect on further extensions of the Max-Combo test in group sequential design. A real-life example will be used for illustration.

#### Impact of censoring and follow-up time and use iwht nonproportional hazards

*Gaohong Dong<sup>1</sup>, Bo Huang<sup>2</sup>, Yu-Wei Chang<sup>3</sup>, Yodit Seifu<sup>4</sup>, ♦James Song<sup>3</sup> and David Hoaglin<sup>5</sup>*

<sup>1</sup>iStats Inc

<sup>2</sup>Pfizer<sup>3</sup>BeiGene<sup>4</sup>Merck<sup>5</sup>U of Massachusetts

james.song@beigene.com

The win ratio has been studied methodologically and applied in data analysis and in designing clinical trials. We distinguish between follow-up time and censoring time, show the impact of censoring on the win ratio, and illustrate the impact of follow-up time. The win ratio can show that benefit by following patients longer, avoiding masking by more frequent but less important outcomes, which occurs in conventional time-to-first-event analyses. For the situation of nonproportional hazards, we demonstrate that the win ratio can be a good alternative to methods such as landmark survival rate, restricted mean survival time, and weighted log-rank tests.

### Beyond Bonferroni Correction: Consistency of Evidence in Clinical Studies

♦ Qian Li, Qiqi Deng<sup>1</sup> and Naitee Ting<sup>1</sup><sup>1</sup>Boehringer Ingelheim Pharmaceuticals

qian.li@bms.com

Multiplicity has been an important topic in clinical trial design, due to multiple endpoints, multiple doses, and multiple interim analyses. Multiple comparison procedures are often applied to control the type I error. Many multiplicity comparison procedures (MCPs) are rooted from Bonferroni Correction with certain variations in forming rejection regions. These MCPs may control the type I error rate strongly, but may not connect the evidence to look at the totality of evidence. On the other hand, methods that do include all evidence may not control the type I error rate strongly. In this paper, we propose a new approach by altering the Bonferroni Correction method, focus on the consistency of evidence, and at the same time control type I error rate strong. The potential application of the new method in clinical trial design and drug evaluation is discussed.

### Session 35: Utilization of RWE in Drug Development: Case Studies

#### Using RWD in Design and Analysis of Clinical Trials - Case Studies

Yanwei Zhang

Takeda Pharmaceutical Company Limited

yanwei.zhang@takeda.com

This presentation will share case studies using real world data (RWD) in design and analysis of clinical trials. In particular, the first case study will focus on the application of RWD in design and analysis of clinical trials at early stages of drug development using Bayesian hierarchical modeling approach. The second case study will focus on the application of RWD for supporting the MAA and HTA submissions in oncology using propensity score matching approach.

#### Actual example of using real world data for drug development

Kiichiro Toyozumi

Janssen Pharmaceutical K.K

ktoyozu@its.jnj.com

Careful consideration is needed when using real world data as historical control data in clinical trial, including but not limited to the source of the data, its quality, comparability of the data, the probability of erroneous conclusion and bias of a treatment effect. Propensity score adjustment is one of the option to address these problems. In this presentation, we will present some example of

clinical trials submitted for regulatory approval which used real world data as historical control data.

### Adjust survival estimates in the presence of treatment switching for HTA

Yuqing Xu<sup>1</sup>, ♦ Meijing Wu<sup>2</sup>, Weili He<sup>2</sup>, Qiming Liao<sup>3</sup> and Yabing Mai<sup>2</sup><sup>1</sup>University of Wisconsin – Madison<sup>2</sup>Abbvie Inc.<sup>3</sup>ViiV Healthcare

meijing.wu@abbvie.com

In oncology clinical trials, characterizing the long-term overall survival (OS) benefit for an experimental drug or treatment regimen (experimental group) is often unobservable if some patients in the control group switch to drugs in the experimental group and/or other cancer treatments after disease progression. A key question often raised by payers and reimbursement agencies is how to estimate the true benefit of the experimental drug group on overall survival that would have been estimated if there were no treatment switches. Several commonly used statistical methods are available to estimate overall survival benefit while adjusting for treatment switching, ranging from naive exclusion or censoring approaches to more advanced methods including inverse probability of censoring weighting (IPCW), iterative parameter estimation (IPE) algorithm or rank-preserving structural failure time models (RPSFTM). However, many clinical trials now have patients switching to different treatment regimens other than the test drugs, and the existing methods cannot handle more complicated scenarios. To address this challenge, we propose two additional methods: stratified RPSFTM and random-forest-based prediction. A simulation study is conducted to assess the properties of the existing methods along with the two newly proposed approaches.

### Real-World Evidence in Regulatory Science

Joan Xie

Seagen

xiejoan904@gmail.com

Under the 21st Century Cures Act the Food and Drug Administration (FDA) is required to develop a framework for a program to evaluate the use of real world evidence (RWE) to help support the approval of a new indication for an approved drug or to satisfy post approval study requirements. Demonstration projects intended to inform the policy development process have already started. This presentation will discuss the types of RWD, infrastructures in FDA for pragmatic trials and examples using the infrastructure and RWD in regulatory science.

### Session 36: Challenges and developments in analyzing complex data

#### Forecasting with clustering based spatially varying auto-regressive model

Sayli Pokal, ♦ Yuzhen Zhou and Trenton Franz

University of Nebraska Lincoln

yuzhenzhou@unl.edu

Crop yield forecasting plays an important role in planning and management of fields. Yet, it becomes especially challenging to do forecasting when we only have short time series. The corn yield data in this study were collected in high spatial resolution (i.e., 10 x 10 m spatial resolution in 800 x 800 m domain) every other year from 2002 to 2016. In this paper, we propose a clustering based spatially varying auto-regressive model for forecasting yield. We compare



the forecasting performance of our model with traditional time series model and a few machine learning algorithms. The results show that for short time series with high spatial resolution data, our proposed model outperforms other models.

#### **Bias-corrected estimation of functional-coefficient autoregressive models with measurement errors**

*Pei Geng*

Illinois State University  
pgeng@ilstu.edu

The functional-coefficient autoregressive models are flexible to fit time series data with covariates. When the time series data is observed with measurement errors, we first investigate the asymptotic bias of the naive estimators. Then we develop a bias-corrected procedure based on the local linear estimation. The asymptotic normality of the bias-corrected estimators is derived. A simulation study shows the efficiency of the proposed method compared to the naive estimation. The proposed approach is also applied to a cybersecurity data.

#### **Bayesian compositional regression with structured priors for microbiome feature selection**

♦*Liangliang Zhang, Yushu Shi, Robert Jenq, Kim-Anh Do and Christine Peterson*

The University of Texas MD Anderson Cancer Center  
lzhang27@mdanderson.org

With the vast development in microbiome studies, researchers has suggested that the human microbiota is becoming a crucial role in understanding health and diseases. However, the applicability of standard variable selection methods are limited by the main challenges of microbiome data. On the one hand, microbiome features are typically quantified as operational taxonomic units (OTUs). They may have a similar impact on the outcome, as they share phylogenetic similarities. On the other hand, the OTU abundances are compositional, since the counts within each sample sum to a constant. To address the challenges posed by these aspects of the data structure, we proposed a Bayesian variable selection model with the following novel features: a generalized transformation to handle the compositional constraint, and a Bayesian prior that encourages the joint selection of microbiome features that are closely related in terms of their genetic sequence similarity. We demonstrate that our proposed method outperforms existing penalized approaches for microbiome variable selection in both simulation and the real data analysis exploring the relationship of the gut microbiome to body mass index (BMI).

### **Session 37: Materials Informatics**

#### **Adaptive exploration and optimization of crystal structures**

♦*Arvind Krishna, Huan Tran, Roshan Joseph and Rampi Ramprasad*

Georgia Institute of Technology  
akrishna39@gatech.edu

We aim to discover new crystal structures through computational methods. The stable crystal structures that exists in the nature have minimum potential energy. Thus, if we can find a structure, that is, its atomic configurations, by minimizing the potential energy, then we might have discovered a new crystal structure. We leverage Density Functional Theory (DFT) to compute the potential energy for a given configuration of the atoms. The problem is challenging because there are infinitely large number of configurations, the DFT code for computing the energy is expensive, and the potential

energy surface is highly non-linear and multi-modal. We propose a novel two-step methodology to find the global minimum. First, we adaptively explore the domain space of all possible configurations of the crystal structure, to generate a representative candidate set of structures. A key feature of our methodology is that we can generate a space-filling design without the knowledge of the boundaries of the domain space. Second, we apply Bayesian optimization over the candidate set to find the global minimum. Gaussian Process modeling along with the Expected Improvement algorithm is used to iteratively update the model and guide the search towards the global minimum. We show the effectiveness of our methodology on toy examples and a real problem.

#### **Accounting for Location Measurement Error in Imaging Data with Application to Atomic Resolution Images of Crystalline Materials**

*Matthew Miller, Matthew Cabral, Elizabeth Dickey, James Lebeau and ♦Brian Reich*

North Carolina State University  
bjreich@ncsu.edu

Scientists use imaging to identify objects of interest and infer properties of these objects. The locations of these objects are often measured with error, which when ignored leads to biased parameter estimates and inflated variance. Current measurement error methods require an estimate or knowledge of the measurement error variance to correct these estimates, which may not be available. Instead, we create a spatial Bayesian hierarchical model that treats the locations as parameters, it using the image itself to incorporate positional uncertainty. We lower the computational burden by approximating the likelihood using a non-contiguous block design around the object locations. We apply this model in a materials science setting to study the relationship between the chemistry and displacement of hundreds of atom columns in crystal structures directly imaged via scanning transmission electron microscopy. Greater knowledge of this relationship can lead to engineering materials with improved properties of interest. We find strong evidence of a negative relationship between atom column displacement and the intensity of neighboring atom columns, which is related to the local chemistry. A simulation study shows our method corrects the bias in the parameter of interest and drastically improves coverage in high noise scenarios compared to non-measurement error models.

#### **Sparse inverse covariance estimation with graph constraints for identifying structures of high entropy alloys**

*Xinrui Liu<sup>1</sup>, Changning Niu<sup>2</sup> and ♦Meng Li<sup>3</sup>*

<sup>1</sup>Shandong Normal University

<sup>2</sup>QuesTek Innovations LLC

<sup>3</sup>Rice University  
meng@rice.edu

There is a wide variety of applications where graph constraints arise naturally from domain knowledge. In this article, we introduce and study a problem of sparse inverse covariance estimation with graph constraints. Building on Tikhonov regularization, we propose to use a penalty decomposition proximal regularization of the Gauss-Seidel algorithm to carry out the estimation. We achieve a closed-form solution with the help of proximity operators. Convergence analysis for the proposed algorithm is provided. We illustrate the proposed methods using simulations and a novel real data application in modeling magnetic moments in high entropy alloys, where the proposed methods lead to interpretable insights in detecting magnetic moment interactions in materials science.

### Session 38: Methodological Advances for Harmonizing Genomics Data to Enable Reproducible Biomedical Research

#### ComBat-seq: batch effect adjustment for RNA-seq count data

Yuqing Zhang<sup>1</sup>, Giovanni Parmigiani<sup>2</sup> and ♦Evan Johnson<sup>3</sup>

<sup>1</sup>Gilead Sciences

<sup>2</sup>Dana Farber Cancer Institute

<sup>3</sup>Boston University  
we.j@bu.edu

The benefit of integrating batches of genomic data to increase statistical power is often hindered by batch effects, or unwanted variation in data caused by differences in technical factors across batches. It is therefore critical to effectively address batch effects in genomic data to overcome these challenges. Many existing methods for batch effects adjustment assume the data follow a continuous, bell-shaped Gaussian distribution. However in RNA-seq studies the data are typically skewed, over-dispersed counts, so this assumption is not appropriate and may lead to erroneous results. Negative binomial regression models have been used previously to better capture the properties of counts. We developed a batch correction method, ComBat-seq, using a negative binomial regression model that retains the integer nature of count data in RNA-seq studies, making the batch adjusted data compatible with common differential expression software packages that require integer counts. We show in realistic simulations that the ComBat-seq adjusted data results in better statistical power and control of false positives in differential expression compared to data adjusted by the other available methods. We further demonstrated in a real data example that ComBat-seq successfully removes batch effects and recovers the biological signal in the data.

#### Improving Predictor Generalizability Using Multiple Studies with Differing Feature Sets

Yujie Wu<sup>1</sup>, Boyu Ren<sup>2</sup>, Giovanni Parmigiani<sup>3</sup> and ♦Prasad Patil<sup>4</sup>

<sup>1</sup>Harvard T.H. Chan School of Public Health

<sup>2</sup>McLean Hospital/Harvard Medical School

<sup>3</sup>Dana-Farber Cancer Institute/Harvard T.H. Chan School of Public Health

<sup>4</sup>Boston University School of Public Health  
prpatil42@gmail.com

Typically, a set of studies or patient cohorts will exhibit heterogeneity in both the marginal distributions of the features used to train a predictor and the conditional distribution of the outcome to predict. In high-dimensional settings, we encounter the additional problem of harmonizing datasets with widely varying levels of overlap in the available features. These discrepancies can lead to poor generalization of a prediction rule learned on a single study and the discarding of potentially useful predictive features that do not fall within convenient overlaps of intersections. We describe progress made in the training of more generalizable prediction functions using multiple studies' worth of data. First, we establish preliminary theoretical guidelines and justification for Cross-Study Learning (CSL), the ensembling of predictors trained in multiple studies. We provide intuition and rules for when to merge studies together and when and how to ensemble single-study predictors. Second, we apply concepts from knowledge transfer in an effort to retain information in non-overlapping feature sets that might otherwise be discarded or ignored out of convenience. We describe conditions under which non-linear and penalized regression can be used to transfer information in the non-overlapping features using functions of the over-

lapping features and improve performance of a CSL predictor.

#### Depth Normalization of Small RNA Sequencing: Using Data and Biology to Select a Best Method

Yannick Duren<sup>1</sup>, Johannes Lederer<sup>1</sup> and ♦Li-Xuan Qin<sup>2</sup>

<sup>1</sup>Ruhr University Bochum

<sup>2</sup>Memorial Sloan Kettering Cancer Center  
qinl@mskcc.org

Deep sequencing has become the most popular tool for transcriptome profiling in cancer research and biomarker studies. Similar to other high through-put profiling technologies such as microarrays, sequencing also suffers from systematic non-biological artifacts that arise from variations in experimental handling. A critical first step in sequencing data analysis is to “normalize” sequencing depth, so that the data can be comparable across the samples. A plethora of analytic methods for depth normalization has been proposed, and different normalization methods may lead to different analysis results with no method found to work systematically best. Currently, it is often up to the data analyst to choose a method based on personal preference and convenience. We developed a data-driven and biology-motivated approach to more objectively guide the selection of a depth normalization method for the data at hand. We assessed the performance of this approach using a unique pair of data sets for the same set of tumor samples that were collected at Memorial Sloan Kettering Cancer Center, and applied it to additional data sets from the Cancer Genome Atlas for further demonstration.

#### A robust normalization method for zero-inflated microbiome sequencing data

Jun Chen

Mayo Clinic  
chen.jun2@mayo.edu

The human microbiome, the collection of microbes associated with the body, has recently received tremendous attention due to its importance in health and disease. Next-generation sequencing technologies pave the way for sequencing-based microbiome studies. Like other types of sequencing data, normalization is the first critical step in microbiome sequencing data analysis used to account for variable library sizes. Current RNA-Seq based normalization methods that have been adapted for microbiome data fail to consider the unique characteristics of microbiome data, which contain a vast number of zeros due to the physical absence or under-sampling of the microbes. Normalization methods that specifically address the zero-inflation remain largely undeveloped. Here we propose Geometric Mean of Pairwise Ratios—a simple but effective normalization method—for zero-inflated sequencing data such as microbiome data. Simulation studies and real datasets analyses demonstrate that the proposed method is more robust than competing methods, leading to more powerful detection of differentially abundant taxa and higher reproducibility of the relative abundances of taxa.

### Session 39: Statistical Method Development Motivated by Biomedical Data Challenges

#### Too many covariates and too few cases? - a comparative study

♦Qingxia Chen, Hui Nian, Yuwei Zhu, Keipp Talbot, Marie Griffin and Frank Harrell

Vanderbilt University Medical Center  
cindy.chen@vumc.org

Prior research indicates that 10-15 cases or controls, whichever fewer, are required per parameter to reliably estimate regression

coefficients in multivariable logistic regression models. This condition may be difficult to meet even in a well-designed study when the number of potential confounders is large, the outcome is rare, and/or interactions are of interest. Various propensity score approaches have been implemented when the exposure is binary. Recent work on shrinkage approaches like lasso were motivated by the critical need to develop methods for the  $p \gg n$  situation, where  $p$  is the number of parameters and  $n$  is the sample size. Those methods, however, have been less frequently used when  $p \approx n$ , and in this situation, there is no guidance on choosing among regular logistic regression models, propensity score methods, and shrinkage approaches. To fill this gap, we conducted extensive simulations mimicking our motivating clinical data, estimating vaccine effectiveness for preventing influenza hospitalizations in the 2011-2012 influenza season. Ridge regression and penalized logistic regression models that penalize all but the coefficient of the exposure may be considered in these types of studies.

#### Quantifying Diagnostic Accuracy Improvement of New Biomarkers for Competing Outcomes

Zheng Wang<sup>1</sup>, Yu Cheng<sup>1</sup>, Eric Seaberg<sup>2</sup> and James Becker<sup>3</sup>

<sup>1</sup>University of Pittsburgh

<sup>2</sup>Johns Hopkins University

<sup>3</sup>University of Pittsburgh  
yucheng@pitt.edu

The net reclassification improvement (NRI) and the integrated discrimination improvement (IDI) were originally proposed to characterize accuracy improvement in predicting a binary outcome, when new biomarkers are added to regression models. These two indices have been extended from dichotomous outcomes to multicategorical and survival outcomes. Working on an AIDS study where the onset of cognitive impairment is competing risks censored by death, we extend the NRI and the IDI to competing risk outcomes, by using cumulative incidence functions to quantify cumulative risks of competing events, and adopting the definitions of the two indices for multi-category outcomes. The "missing" category due to independent censoring is handled through inverse probability weighting. Various competing risks models are considered, such as the Fine and Gray, multistate, and multinomial logistic models. Estimation methods for the NRI and the IDI from competing risks data are presented. The inference for the NRI is constructed based on asymptotic normality of its estimator, and the bias-corrected and accelerated bootstrap procedure is applied for the IDI inference. Simulations demonstrate that the proposed inferential procedures perform very well. The Multicenter AIDS Cohort Study is used to illustrate the practical utility of the extended NRI and IDI for competing risks outcomes.

#### Analysis of Generalized Semiparametric Mixed Varying-Coefficients Models for Longitudinal Data

Yanqing Sun<sup>1</sup>, Li Qi<sup>2</sup>, Fei Heng<sup>3</sup> and Peter Gilbert<sup>4</sup>

<sup>1</sup>University of North Carolina at Charlotte

<sup>2</sup>Sanofi, Bridgewater, U.S.A.

<sup>3</sup>University of North Florida

<sup>4</sup>University of Washington and Fred Hutchinson Cancer Research Center

yasun@uncc.edu

The generalized semiparametric mixed varying-coefficient effects model for longitudinal data can accommodate a variety of link functions and flexibly model different types of covariate effects, including time-constant, time-varying, and covariate-varying effects. The time-varying effects are unspecified functions of time and the

covariate-varying effects are nonparametric functions of a possibly time-dependent exposure variable. A semiparametric estimation procedure is developed that uses local linear smoothing and profile weighted least squares, which requires smoothing in the two different and yet connected domains of time and the time-dependent exposure variable. The asymptotic properties of the estimators of both nonparametric and parametric effects are investigated. In addition, hypothesis testing procedures are developed to examine the covariate effects. The finite-sample properties of the proposed estimators and testing procedures are examined through simulations, indicating satisfactory performances. The proposed methods are applied to analyze the ACTG 244 clinical trial to investigate the effects of antiretroviral treatment switching in HIV infected patients before and after developing the T215Y antiretroviral drug resistance mutation.

#### Explained Variance Decompositions for Mediation Effect Sizes with Multiple Exposures

Shanshan Zhao<sup>1</sup>, Yue Jiang<sup>2</sup> and Jason Fine<sup>3</sup>

<sup>1</sup>NIEHS/NIH

<sup>2</sup>Duke University

<sup>3</sup>University of North Carolina - Chapel Hill  
shanshan.zhao@nih.gov

Mediation analysis assesses relative contributions of direct and indirect effects to explore underlying processes by which exposures affect outcomes. Although methods to test for mediation effects have been widely used, relatively little research has focused on developing effect size measures, especially in settings with multiple exposures where standard use of proportion mediated is not applicable. We propose an effect size measure based on variance decompositions which summarize the overall effect of multiple exposures. We further develop multistage constrained least squares estimators to effectively constrain the range of the proposed measure in the range of 0 to 1, a desirable feature for relative effect size measures. Closed forms and asymptotic properties are derived and evaluated through numerical simulation. We apply the proposed measures to the Agricultural Health Study to explore the effect of multiple smoking related exposures on lung functions through methylation at various CpG sites.

#### Session 40: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences

##### Measuring Diagnostic Accuracy and Selecting Optimal Cut-points for K-class Diseases Based on Concordance and Discordance with Application

Jing Kersey, Hani Samawi, Jingjing Yin, Haresh Rochani and Xinyan Zhang

Georgia Southern University  
jk01810@georgiasouthern.edu

An essential aspect of medical diagnostic testing using biomarkers is to find an optimal cut-point that categorizes a patient as diseased or healthy. This aspect can be extended to the diseases which can be classified into more than two classes. For diseases with general  $k$  ( $k > 2$ ) classes, well-established measures include hypervolume under the manifold and the generalized Youden Index. Another two diagnostic accuracy measures, maximum absolute determinant (MADET) and Kullback-Leibler divergence measure (KL) are recently proposed. This research proposes a new measure of diagnostic accuracy based on concordance and discordance (CD) for diseases with  $k$  ( $k > 2$ ) classes and uses it as a cut-points selection

criterion. The CD measure utilizes all the classification information and provides more balanced class probabilities. Power studies and simulations show that the optimal cut-points selected with CD measure may be more accurate for early-stage detection in some scenarios compared with other available measures. As well, an example of an actual dataset from the medical field will be provided using the proposed CD measure.

### Pathway-Structured Predictive Modeling for Multi-Level Drug Response in Multiple Myeloma

♦ *Xinyan Zhang<sup>1</sup>, Wenzhuo Zhuang<sup>2</sup> and Nengjun Yi<sup>3</sup>*

<sup>1</sup>Kennesaw State University

<sup>2</sup>Soochow University

<sup>3</sup>University of Alabama at Birmingham

xzhang47@kennesaw.edu

Motivation: Molecular analyses suggest that myeloma is composed of distinct subtypes that have different molecular pathologies and various response rates to certain treatments. Drug responses in multiple myeloma (MM) are usually recorded as a multi-level ordinal outcome. One of the goals of drug response studies is to predict which response category any patients belong to with high probability based on their clinical and molecular features. However, as most of genes have small effects, gene-based models may provide limited predictive accuracy. In that case, methods for predicting multi-level ordinal drug responses by incorporating biological pathways are desired but have not been developed yet. Results: We propose a pathway-structured method for predicting multi-level ordinal responses using a two-stage approach. We first develop hierarchical ordinal logistic models and an efficient quasi-Newton algorithm for jointly analyzing numerous correlated variables. Our two-stage approach first obtains the linear predictor (called the pathway score) for each pathway by fitting all predictors within each pathway using the hierarchical ordinal logistic approach, and then combines the pathway scores as new predictors to build a predictive model. We applied the proposed method to two publicly available datasets for predicting multi-level ordinal drug responses in MM using large-scale gene expression data and pathway information. Our results show that our approach not only significantly improved the predictive performance compared with the corresponding gene-based model but also allowed us to identify biologically relevant pathways.

### Application of Empirical Likelihood methods on bivariate Mean Residual Life function

♦ *Ali Jinnah and Yichuan Zhao*

Georgia State University

alijinnah1234@outlook.com

Kulkarni and Rattihalli (2002) proposed an estimator for the bivariate mean residual life (MRL) function. In this paper, we apply the empirical likelihood (EL) and adjusted empirical likelihood (AEL) methods to the MRL function. The Wilk's theorem is established under general conditions. We profile the nuisance parameter in the EL and develop EL for the univariate MRL function. Extensive simulation studies show EL methods for both bivariate and one-dimensional MRL functions perform better than the normal approximation (NA) method in terms of coverage probabilities. AEL methods result in noticeable better coverage probability. AEL method based on F-distribution calibration results in better coverage probability for small sample sizes. Two real data sets are used to illustrate the proposed procedure.

### Generalized mean residual life models for case-cohort and

### nested case-control studies

♦ *Peng Jin, Anne Zeleniuch-Jacquotte and Mengling Liu*

New York University School of Medicine

peng.jin@nyulangone.org

Mean residual life (MRL) is the remaining life expectancy of a subject who has survived to a certain time point and can be used as an alternative to hazard function for characterizing the distribution of a time-to-event variable. Inference and application of MRL models have primarily focused on full-cohort studies. In practice, case-cohort and nested case-control designs have been commonly used within large cohorts that have long follow-up and study rare diseases, particularly when studying costly molecular biomarkers. They enable prospective inference as the full-cohort design with significant cost-saving benefits. In this paper, we study the modeling and inference of a family of generalized MRL models under case-cohort and nested case-control designs. Built upon the idea of inverse selection probability, the weighted estimating equations are constructed to estimate regression parameters and baseline MRL function. Asymptotic properties of the proposed estimators are established and finite-sample performance is evaluated by extensive numerical simulations. An application to the New York University Women's Health Study is presented to illustrate the proposed models and demonstrate a model diagnostic method to guide practical implementation.

## Session 41: Statistical methods for complex human genetic data

### A Kernel-Based Neural Network for High-dimensional Genetic Data Analysis

♦ *Qing Lu, Xiaoxi Shen and Xiaoran Tong*

University of Florida

lucienq@php.ufl.edu

Artificial intelligence (AI) is a thriving research field with many successful applications in areas such as computer vision and speech recognition. Neural-network-based methods (e.g., deep learning) play a central role in modern AI technology. While neural-network-based methods also hold great promise for genetic research, the high-dimensionality of genetic data, the massive amounts of study samples, and complex relationships between genetic variants and disease outcomes bring tremendous analytic and computational challenges. To address these challenges, we propose a kernel-based neural network (KNN) method. KNN inherits features from both linear mixed models (LMM) and classical neural networks and is designed for high-dimensional genetic data analysis. Unlike the classic neural network, KNN summarizes a large number of genetic variants into kernel matrices and uses the kernel matrices as input matrices. Based on the kernel matrices, KNN builds a feedforward neural network to model the complex relationship between genetic variants and a disease outcome. Minimum norm quadratic unbiased estimation and batch training are implemented in KNN to accelerate the computation, making KNN applicable to massive datasets with millions of samples. Through simulations, we demonstrate the advantages of KNN over LMM in terms of prediction accuracy and computational efficiency. We also apply KNN to the large-scale UK Biobank dataset, evaluating the role of a large number of genetic variants on multiple complex diseases.

### A Bayesian Graphical model to delineate essential enhancer regulations using genome editing data

*Hao Wang*

Michigan State University  
wangha73@msu.edu

Genome-wide association study (GWAS) was proposed as a powerful tool to associate genotypes with phenotypes and has been used to discover millions of disease-associated SNPs. However, the pervasive spatial correlations between those disease-associated SNPs hamper the ability to discover causal SNPs and blur the underlying mechanisms of GWAS SNPs. Recently, advanced genome-editing technology CRISPR has been developed, which enables people to mutate the genome. This technique establishes a mapping between genotype changes and phenotype changes directly, thus providing the causal information and a revolutionary opportunity to boost causal SNP identification. But, the complex structures of genomes, i.e. unknown biological regulatory grammars, makes the CRISPR signal noisy and hard to interpret. It is therefore important to develop a new statistical approach to decode the unknown mixtures of CRISPR signals. In this talk, I will introduce our statistical generative model to decode the CRISPR data by integrating domain-knowledge and the new mechanistic insights on GWAS SNPs. By applying the model to a real-world CRISPR dataset, I will demonstrate the accuracy and advantages of our models, and further discuss its potential in prioritizing causal GWAS SNPs.

#### Statistical inference of 3D genome structures and its applications in human genetics

Jianrong Wang

Michigan State University  
wangj164@msu.edu

Advances in population genetics, such as genome-wide association studies (GWAS), have made substantial progress to catalog which individual genetic variants are associated with specific human diseases. But we still do not know how these genetic variants, especially non-coding variants, lead to observed phenotypes. Therefore, there is a significant need to elucidate the underlying mechanisms by which the effects of individual genetic variants propagate to organism phenotypes through intermediate molecular disruptions. In recent years, high-throughput 3D chromatin contact maps have been generated for diverse tissues or cell-types, providing new information to link non-coding genetic variants to target genes and associated pathways. Due to the 'big data' challenges, efficient and robust statistical models and machine learning algorithms are needed to integrate 3D chromatin structure information for improved GWAS interpretations. I will discuss several models we developed on this topic, and will demonstrate how these methods can be leveraged to improve the statistical power and identify potential causal regulatory SNPs and pathways in human diseases. The derived network-level predictions will provide insights on gene regulation, chromatin architecture and disease mechanisms, leading to novel genomics-based diagnostics and therapeutics.

#### Multivariate partial linear varying coefficients model for genetic association studies with multiple longitudinal traits

♦Honglang Wang<sup>1</sup>, Jingyi Zhang<sup>2</sup> and Yuehua Cui<sup>3</sup>

<sup>1</sup>Indiana University-Purdue University Indianapolis

<sup>2</sup>Wells Fargo, Charlotte

<sup>3</sup>Michigan State University  
hlwang@iupui.edu

Genetic pleiotropy refers to the situation in which a gene can influence multiple traits. For some complex diseases, multiple phenotypic measurements can be used to quantify the disease status. Such correlated phenotypes often share common genetic determinants. For multivariate longitudinal data, when multiple response

variables are jointly measured over time, the correlation information between multivariate longitudinal responses can be taken into account to identify any pleiotropic effects. In this work, we proposed a multivariate partially linear varying coefficients model to identify genetic variants with their effects potentially modified by environmental factors or varying over time. We derived a testing framework to jointly test the association of genetic factors with a bivariate phenotypic trait while taking the varying genetic effects into account. We extended the quadratic inference functions to deal with the longitudinal correlations and used penalized splines for the approximation of nonparametric coefficients. Theoretical results such as consistency and asymptotic normality of the estimates are established. The performance of the testing procedure was evaluated through Monte Carlo simulation studies. The utility of the method was demonstrated with a real data set from a pain sensitivity study, in which SNPs associated with systolic blood pressure (SBP) and diastolic blood pressure (DBP) were identified. These SNP effects on blood pressure show a nonlinear relationship as the dosage level of Donutamine varies, indicating the potential for personalized treatment.

#### Session 42: Recent advances in statistical methods for missing data, measurement error and biased sampling

##### A New Bayesian Joint Model for Longitudinal Count Data with Many Zeros, Intermittent Missingness, and Dropout with Applications to HIV Prevention Trials

Jing Wu<sup>1</sup>, ♦Ming-Hui Chen<sup>2</sup>, Elizabeth Schifano<sup>2</sup>, Joseph Ibrahim<sup>3</sup> and Jeffrey Fisher<sup>2</sup>

<sup>1</sup>University of Rhode Island

<sup>2</sup>University of Connecticut

<sup>3</sup>University of North Carolina at Chapel Hill  
ming-hui.chen@uconn.edu

In longitudinal clinical trials, it is common that subjects may permanently withdraw from the study (dropout), or return to the study after missing one or more visits (intermittent missingness). It is also routinely encountered in HIV prevention clinical trials that there is a large proportion of zeros in count response data. In this paper, a sequential multinomial model is adopted for dropout and subsequently a conditional model is constructed for intermittent missingness. The new model captures the complex structure of missingness and incorporates dropout and intermittent missingness simultaneously. The model also allows us to easily compute the predictive probabilities of different missing data patterns. A zero inflated Poisson mixed-effects regression model is assumed for the longitudinal count response data. We also propose an approach to assess the overall treatment effects under the zero-inflated Poisson model. We further show that the joint posterior distribution is improper if uniform priors are specified for the regression coefficients under the proposed model. Variations of the g-prior, Jeffreys prior, and maximally dispersed normal prior are thus established as remedies for the improper posterior distribution. An efficient Gibbs sampling algorithm is developed using a hierarchical centering technique. A modified logarithm of the pseudomarginal likelihood (LPML) and a concordance based area under the curve (AUC) criterion are used to compare the models under different missing data mechanisms. We then conduct an extensive simulation study to investigate the empirical performance of the proposed methods, and further illustrate the methods using real data from an HIV prevention clinical trial.

**Bayesian nonparametrics for missing data in EHRs***Michael Daniels*University of Florida  
daniels@ufl.edu

We propose a framework for missing data and causal inference in EHRs based on Bayesian nonparametric models for the distribution of the observed data. To then identify and estimate quantities of interest, (observed data) uncheckable assumptions are required. A natural way to incorporate uncertainty in these assumptions is by introduction of sensitivity parameters (within these assumptions) that are given informative priors. We demonstrate the ease and power of the proposed framework for causal inference in the presence of missing data using EHRs.

**Semiparametric Generalized Linear Models for Analysis of Longitudinal Data with Biased Observation-level Sampling**♦ *Paul Rathouz<sup>1</sup> and Jacob Maronge<sup>2</sup>*<sup>1</sup>University of Texas at Austin<sup>2</sup>University of Wisconsin-Madison  
paul.rathouz@austin.utexas.edu

Abstract: Rathouz and Gao (2009) proposed a novel class of generalized linear models indexed by a linear predictor and a link function for the mean of  $(Y|X)$ . In this class, the distribution of  $(Y|X)$  is left unspecified and estimated from the data via exponential tilting of a reference distribution, yielding a semiparametric response model that is a member of the natural exponential family. We have since developed generalized case-control sampling designs for univariate data in this class of models. In this talk, we show how these designs extend to longitudinal studies with time-point specific sampling plans, where sampling may depend on earlier observations and/or on an auxiliary variable. KW: Case-control, exponential tilting, generalized linear models, longitudinal data, marginal models, outcome dependent sampling, quasilielihood.

**Variable Selection for Proportional Hazards Models with High Dimensional Covariates subject to Measurement Error**♦ *Baojiang Chen<sup>1</sup>, Ao Yuan<sup>2</sup> and Grace Yi<sup>3</sup>*<sup>1</sup>The University of Texas Health Science Center at Houston<sup>2</sup>Georgetown University<sup>3</sup>University of Western Ontario  
baojiang.chen@uth.tmc.edu

Methods of analyzing high dimensional data are often challenged by the presence of measurement error in variables, a common issue arising from various applications. Conducting Naïve analysis with measurement error effects ignored usually gives biased results. However, relatively little research has been focused on this topic. In this paper, we consider this important problem and discuss variable selection for proportional hazards models with high dimensional covariates subject to measurement error. We propose a penalized “corrected” likelihood-based method to simultaneously address the measurement error effects and perform variable selection. We establish theoretical results including the consistency, the oracle property, and the asymptotic distribution of the proposed estimator. Simulation studies are conducted to assess the finite sample performance of the proposed method. To illustrate the use of our method, we apply the proposed method to analyze a data set arising from the breast cancer study.

**Session 43: Statistical Learning Advancement for Inference with Complex Biomedical Data****Peel Learning for Pathway-related Outcome Prediction**♦ *Rui Feng<sup>1</sup>, Yuantong Li<sup>2</sup>, Mengying Yan<sup>3</sup> and Edward Cantu<sup>1</sup>*<sup>1</sup>University of Pennsylvania<sup>2</sup>Purdue University<sup>3</sup>George Washington University  
rui.feng@upenn.edu

Traditional regression methods have limited capability of predicting a clinical outcome using a set of genes that genes may correlate, interact, and jointly affect the outcome. Current deep learning methods may provide better predictions, but they often need to be trained on a large amount of data. Gene expression studies often have limited sample sizes. In this paper, we proposed peel learning, a deep learning framework that incorporates the known relationship among genes. In each layer of learning, we focused on multiple local substructures trimmed from the parent features. We derived a computational-efficient peeling algorithm where features are decomposed into independent components and then summarized within each substructure. The substructures are gradually reduced in size over layers and the parameters in layers are optimized through backpropagation. We evaluated the performance of our method through simulations. We applied the proposed method to predict lung transplantation outcome using the gene expression profiles in donors’ lungs. Our method showed improved prediction accuracy, especially in small data, compared to conventional penalized regression, classification trees, feed-forward neural network, and a neural network assuming fully connected pathways.

**Using Statistical Learning to Promote Evidence-based Precision Health Care***Lu Wang*University of Michigan  
luwang@umich.edu

In this talk, we present recent advances and statistical developments for evaluating Dynamic Treatment Regimes (DTR), which allow the treatment to be dynamically tailored according to evolving subject-level data. Identification of an optimal DTR is a key component for precision medicine and personalized health care. Specific topics covered in this talk include several recent projects with robust and flexible methods developed for the above research area. We will first introduce a dynamic statistical learning method, adaptive contrast weighted learning (ACWL), which combines doubly robust semiparametric regression estimators with flexible machine learning methods. We will further develop a tree-based reinforcement learning (T-RL) method, which builds an unsupervised decision tree that maintains the nature of batch-mode reinforcement learning. Unlike ACWL, T-RL handles the optimization problem with multiple treatment comparisons directly through a purity measure constructed with augmented inverse probability weighted estimators. T-RL is robust, efficient and easy to interpret for the identification of optimal DTRs. However, ACWL seems more robust against tree-type misspecification than T-RL when the true optimal DTR is non-tree-type. At the end of this talk, we will also present a new Stochastic-Tree Search method called ST-RL for evaluating optimal DTRs.

**Multi-omic integration to reveal functional consequences of DNA alterations in tumor**♦ *Xiaoyu Song<sup>1</sup>, Jiayi Ji<sup>1</sup>, Lin Chen<sup>2</sup> and Pei Wang<sup>1</sup>*<sup>1</sup>Icahn School of Medicine at Mount Sinai

<sup>2</sup>University of Chicago  
xiaoyu.song@mountsinai.org

We proposed an integrative analysis tool iProFun to screen for DNA alterations perturbing proteogenomic functional traits. Specifically, we considered multiple types of DNA alterations including mutation, DNA methylation and copy number variation (CNV) and analyzed their impacts on multiple molecular quantitative traits simultaneously including mRNA, protein, and phosphoprotein levels. We aim to identify genes whose DNA alterations have cis-associations with either some or all omic traits. In comparison with analyzing each molecular trait separately, the joint modeling of multi-omics data enjoys enhanced power and it also achieves better accuracy in inferring cis-associations unique to certain type(s) of molecular trait(s). We applied iProFun to multiple cancer types from Clinical Proteomic Tumor Analysis Consortium (CPTAC), and identified DNA alterations with preserving effects through transcriptional, translational, and post-translational levels, and prioritized gene targets for tumor initiation and progression.

#### Microbial Network Recovery by Compositional Graphical Lasso

Chuan Tian, Duo Jiang, Thomas Sharpton and <sup>◆</sup>Yuan Jiang  
Oregon State University  
yuan.jiang@stat.oregonstate.edu

Network models such as graphical models have become a useful approach to studying the interactions between microbial taxa given the microbiome data deluge. Recently, 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 count nature of microbiome data, where abundances of microbial taxa are not directly measured but are presented by error-prone counts. Adding to the challenge is that the sum of the counts within each sample, termed “sequencing depth”, can vary drastically across samples. To address these issues, we adopt a logistic normal multinomial model explicitly incorporating the sequencing depth and develop an algorithm that iterates between Newton-Raphson and graphical lasso for model estimation. We call this new approach “compositional graphical lasso”. We have established the convergence of the algorithm. Additionally, we illustrate the advantage of compositional graphical lasso in comparison to current methods under a variety of simulation scenarios and also demonstrate the applicability of compositional graphical lasso to a real microbiome data set.

#### Session 44: New Developments in High-Dimensional Data Analysis

##### Sufficient dimension folding in regression via distance covariance for matrix-valued predictors

<sup>◆</sup>Wenhui Sheng<sup>1</sup> and Qingcong Yuan<sup>2</sup>  
<sup>1</sup>Marquette University

<sup>2</sup>Miami University  
wenhui.sheng@marquette.edu

In modern data, when predictors are matrix/array-valued, building a reasonable model is much more difficult due to the complicate structure. However, dimension folding that reduces the predictor dimensions while keeps its structure is critical in helping to build a useful model. In this paper, we develop a new sufficient dimension folding method using distance covariance for regression in such a case. The method works efficiently without strict assumptions on the predictors. It is model-free and nonparametric, but neither smoothing

techniques nor selection of tuning parameters is needed. Moreover, it works for both univariate and multivariate response cases. In addition, we propose a new method of local search to estimate the structural dimensions. Simulations and real data analysis support the efficiency and effectiveness of the proposed method.

##### Multivariate Dimension Reduction and the Dual Central Subspaces

<sup>◆</sup>Ross Iaci<sup>1</sup>, Xiangrong Yin<sup>2</sup> and Lixing Zhu<sup>3</sup>

<sup>1</sup>The College of William and Mary

<sup>2</sup>University of Kentucky

<sup>3</sup>Hong Kong Baptist University  
riaci@wm.edu

Existing dimension reduction methods in multivariate analysis have focused on reducing sets of random vectors into equivalently sized dimensions, while methods in regression settings have largely focused on decreasing the dimension of the predictor variables. However, for problems involving a multivariate response, reducing the dimension of the response vector is also desirable and important. In this talk, a nonparametric method that provides a dimension reduction of two multivariate random vectors without requiring the dimensions of the reduction to be equal is discussed. This method is also applicable for random vectors labeled as predictor and response and thus, provides a powerful tool for dimension reduction in a multivariate regression setting. To this end, a new concept termed the Dual Central Subspaces (DCS) is introduced, where the estimation of these subspaces provides a sufficient dimension reduction of the random vectors. To recover the DCS, a higher-order information measure based on the Kullback-Leibler (KL) divergence is used, rather than extending traditional methods for estimating the Central Subspace (CS) that recover information from moments, such as SIR and SAVE. Using this information based measure enables the recovery of both linear and nonlinear relationships that exist between random vectors and thereby, allows for a more complete identification of the DCS while treating both vectors equivalently. To achieve a dimension reduction, a bootstrap procedure to estimate the dimensions of the DCS will be discussed. The method is illustrated in a real-world example by analyzing an environmental dataset from Los Angeles County in order to study the associations that exist between mortality and environmental conditions.

##### Generalized Spatially Varying Coefficient Models

<sup>◆</sup>Myungjin Kim and Li Wang

Iowa State University  
mjkim@iastate.edu

In this paper, we introduce a new class of nonparametric regression models, called generalized spatially varying coefficient models (GSVCMs), for data distributed over complex domains. For model estimation, we propose a nonparametric quasi-likelihood approach using the bivariate penalized spline approximation technique. We show that our estimation procedure is able to handle irregularly-shaped spatial domains with complex boundaries. Under some regularity conditions, the estimator for the coefficient function is proved to be consistent in the L<sub>2</sub> sense and its convergence rate is established. We develop a numerically stable algorithm using penalized iteratively reweighted least squares method to estimate the coefficient functions in GSVCMs. To gain efficiency in the computation for large-scale data, we further propose a QR decomposition-based algorithm, which requires only sub-blocks of the design matrix to be computed at a time, so that it allows efficient estimation of GSVCMs for large datasets with modest computer hardware. The finite sample performance of the GSVCM and its estimation method

is examined by simulations studies. The proposed method is also illustrated by an analysis of the crash data in Florida.

### Spatial Autoregressive Partially Linear Varying Coefficient Models

Jingru Mu

Kansas State University  
jingrumu@ksu.edu

In this article, we consider a class of partially linear spatially varying coefficient autoregressive models for data distributed over complex domains. We propose approximating the varying coefficient functions via bivariate splines over triangulation to deal with the complex boundary of the spatial domain. Under some regularity conditions, the estimated constant coefficients are asymptotically normally distributed, and the estimated varying coefficients are consistent and possess the optimal convergence rate. A penalized bivariate spline estimation method with a more flexible choice of triangulation is proposed. We further develop a fast algorithm to calculate the geodesic distance. The proposed method is much more computationally efficient than the local smoothing methods, and thus capable of handling large scales of spatial data. In addition, we propose a model selection approach to identify predictors with constant and varying effects. The performance of the proposed method is evaluated by simulation examples and the Sydney real estate dataset.

### Session 45: Empirical Likelihood Methods and Bayesian Variable Selection

#### Bayesian empirical likelihood based methods

♦ Yichen Cheng and Yichuan Zhao

Georgia State University  
ycheng11@gsu.edu

Empirical likelihood is a very powerful nonparametric tool that does not require any distributional assumptions. In the talk, we will talk about two related projects based on Bayesian empirical likelihood. Lazar (2003) showed that if you replace the usual likelihood component in the Bayesian posterior likelihood with the empirical likelihood, then posterior inference is still valid when the functional of interest is a smooth function of the posterior mean. However, it is not clear whether similar conclusions can be obtained for parameters defined in terms of U-statistics. In the first project, we answer this question using Jackknife empirical likelihood. In the second project, we explore the possibility of constructing Bayesian empirical likelihood for variable selection.

#### The Bayesian Elastic Net based on Empirical Likelihood

Adel Bedoui<sup>1</sup> and ♦ Chul Moon<sup>2</sup>

<sup>1</sup>Boehringer Ingelheim

<sup>2</sup>Southern Methodist University  
chulm@mail.smu.edu

The elastic net estimates can be interpreted as Bayesian posterior estimates when the regression parameters have a prior that compromises between Gaussian and independent Laplace (i.e., double-exponential) priors. A significant challenge in the elastic net is that it assumes that data are normally distributed, which makes it not robust to model misspecification. In this article, we propose a Bayesian semiparametric approach for an elastic net model that is based on empirical likelihood. This approach relaxes the normal assumption on data, and hence we avoid problems with model misspecification. Under the Bayesian empirical likelihood approach,

the resulting posterior distribution lacks a closed-form and has non-convex support, which makes the implementation of traditional Markov chain Monte Carlo methods such as Gibbs sampling and Metropolis-Hastings very challenging. To solve the nonconvex optimization and nonconvergence problems, we implement the Hamiltonian Monte Carlo approach.

#### Reduce the computation in jackknife empirical likelihood for comparing two correlated Gini indices

Kangni Alemjrodo and ♦ Yichuan Zhao

Georgia State University  
yichuan@gsu.edu

The Gini index has been widely used as a measure of income (or wealth) inequality in social sciences. To construct a confidence interval for the difference of two Gini indices from the paired samples, Wang and Zhao (2016) used a profile jackknife empirical likelihood after maximization over a nuisance parameter and established Wilks' theorem. However, profiling could be very expensive. In this paper, we propose an alternative approach of the jackknife empirical likelihood method to reduce the computational cost. We also investigate the adjusted jackknife empirical likelihood and the bootstrap-calibrated jackknife empirical likelihood to improve coverage accuracy for small samples. Simulations show that the proposed methods perform better than Wang and Zhao's methods in terms of coverage accuracy and computational time. Two real data applications proved that the proposed methods work perfectly in practice.

#### Full likelihood inference for abundance from capture-recapture data

Pengfei Li

University of Waterloo  
pengfei.li@uwaterloo.ca

Capture-recapture experiments are widely used to collect data needed to estimate the abundance of a closed population. To account for heterogeneity in the capture probabilities, Huggins (1989) and Alho (1990) proposed a semiparametric model in which the capture probabilities are modelled parametrically and the distribution of individual characteristics is left unspecified. A conditional likelihood method was then proposed to obtain point estimates and Wald-type confidence intervals for the abundance. Empirical studies show that the small-sample distribution of the maximum conditional likelihood estimator is strongly skewed to the right, which may produce Wald-type confidence intervals with lower limits that are less than the number of captured individuals or even negative. In this talk, we present a full empirical likelihood approach based on this model. We show that the null distribution of the empirical likelihood ratio for the abundance is asymptotically chi-square with one degree of freedom, and the maximum empirical likelihood estimator achieves semiparametric efficiency. Simulation studies show that the empirical-likelihood-based method is superior to the conditional-likelihood-based method: its confidence interval has much better coverage, and the maximum empirical likelihood estimator has a smaller mean square error. We analyze three data sets to illustrate its advantages.

### Session 46: Statistical Process Control and Detection of Change-Point

#### Variance Change Point Detection Under a Smoothly-Changing Mean Trend with Application to Liver Procurement

♦ Zhenguang Gao<sup>1</sup>, Zuofeng Shang<sup>2</sup>, Pang Du<sup>3</sup> and John Robertson<sup>3</sup>

<sup>1</sup>School of Mathematical Sciences, Shanghai Jiao Tong University



<sup>2</sup>IUPUI<sup>3</sup>Virginia Tech

gaozhenguo3@126.com

Literature on change point analysis mostly requires a sudden change in the data distribution, either in a few parameters or the distribution as a whole. We are interested in the scenario, where the variance of data may make a significant jump while the mean changes in a smooth fashion. The motivation is a liver procurement experiment monitoring organ surface temperature. Blindly applying the existing methods to the example can yield erroneous change point estimates since the smoothly changing mean violates the sudden-change assumption. We propose a penalized weighted least-squares approach with an iterative estimation procedure that integrates variance change point detection and smooth mean function estimation. The procedure starts with a consistent initial mean estimate ignoring the variance heterogeneity. Given the variance components the mean function is estimated by smoothing splines as the minimizer of the penalized weighted least squares. Given the mean function, we propose a likelihood ratio test statistic for identifying the variance change point. The null distribution of the test statistic is derived together with the rates of convergence of all the parameter estimates. Simulations show excellent performance of the proposed method. Application analysis offers numerical support to non invasive organ viability assessment by surface temperature monitoring. Supplementary materials for this article are available online.

#### Optimal rate of convergence of multivariate nonparametric change point detection

Xin Xing<sup>1</sup>, Zuofeng Shang<sup>2</sup>, ♦Pang Du<sup>3</sup>, Hongyu Miao<sup>4</sup> and Jun Liu<sup>1</sup>

<sup>1</sup>Harvard University<sup>2</sup>New Jersey Institute of Technology<sup>3</sup>Virginia Tech<sup>4</sup>The University of Texas Health Science Center at Houston  
pangdu@vt.edu

Change-point analysis of an unlabeled sample of observations consists in, first, testing whether a change in the distribution occurs within the sample, and second, if a change occurs, estimating the change-point instant after which the distribution of the observations switches from one distribution to another different distribution. Recently, the nonparametric testing are popular to serve the first purpose. However, there is still limited work on studying the convergence rate of the change point since the estimation is usually involved a infinite series of testing statistics. In this paper, we establish a non-asymptotic theory for nonparametric density estimation in a reproducing kernel Hilbert space. Based on the derived non-asymptotic bound, we are able to derive the convergence rate of the proposed change point estimator.

#### A Method of Optimizing the Control Charts for Finite Sequence of Observations

♦Dong Han<sup>1</sup>, Fugee Tsung<sup>2</sup> and Jinguo Xian<sup>1</sup>

<sup>1</sup>Shanghai Jiao Tong University<sup>2</sup>Hong Kong University of science and technology  
donghan@sjtu.edu.cn

We propose a method of optimizing the control limit for optimizing the control charts with the given charting statistics to detect a change in distribution of finite sequence of observations. The optimized control chart is proved to have the smallest average value of some kind of detection delay among all control charts with the false alarm rate no less than a given value. The method is illustrated by numerical simulations of the three optimized control charts, Shewhart, EWMA and CUSUM charts, in detecting mean shifts.

#### Fault Classification for High-dimensional Data Streams: A Directional Diagnostic Framework Based on Multiple Hypothesis Testing

Dongdong Xiang

East China Normal University

terryxdd@163.com

In various modern statistical process control applications that involve high-dimensional data streams (HDDS), accurate fault diagnosis of out-of-control (OC) data streams is becoming crucial. The existing diagnostic approaches either focus on moderatedimensional processes or are unable to determine the shift direction accurately, especially when the signal-to-noise ratio is low. In this paper, we conduct a bold trial and consider the fault classification problem of the mean vector of HDDS where determining the shift direction of the OC data streams is important to perform customized repairs. To this end, under the basic assumptions that the in-control data streams are normal with mean 0 and variance 1, and that the high-dimensional observations after the alarm are solely OC, the problem is formulated into a three-classification multiple testing framework, and an efficient data-driven diagnostic procedure is developed to minimize the expected number of false positives while controlling the missed discovery rate at satisfactory level. The procedure is statistically optimal and computationally efficient, and improves the diagnostic effectiveness by taking into account directional information, which provides insights to guide further decisions. Both theoretical and numerical results reveal the superiority of the new method.

#### Session 47: Novel computational techniques for analyzing large scale biostatistical data

##### Reduction of Bias Due to Misclassified Exposures using Instrumental Variables

Christopher Manuel, ♦Samiran Sinha and Suojin Wang

Texas A&amp;M University

sinha@stat.tamu.edu

Exposure variables are often misclassified in observational studies. Any analysis that does not make proper adjustments for misclassification may result in biased estimates of model parameters and that may lead to distorted inferences. The case where a multicategory exposure variable having more than two nominal categories, or the analysis when no validation data are available to assess the misclassification probabilities, is seldom considered in the literature. In this talk I will present a novel method of analyzing cohort data with a misclassified multicategory exposure variable with the help of instrumental variables in lieu of a validation dataset. For the parameter estimation, a variational Bayesian inference procedure aided by the automatic differentiation variational inference technique is used. Operating characteristics of the method are assessed and compared with existing approaches through simulation studies. I will present the simulation results and an illustrating example on the US breast cancer mortality data sampled from the Surveillance Epidemiology and End Results database.

##### Bayesian nonparametric bi-clustering of microbiome data

Yang Ni

Texas A&amp;M University

yni@stat.tamu.edu

We develop a novel Bayesian nonparametric bi-clustering algorithm for microbiome data. We propose a mixture model framework to dynamically dichotomize multinomial data into two categories. On top

of the mixture layer, a double feature allocation model is imposed on the binary mixture indicators. Double feature allocation model clusters both observations and variables (OTUs). Moreover, it allows for overlapping clustering structures. When prior information regarding the clustering structure is available, it is straightforward to incorporate it in the proposed Bayesian method. We demonstrate the utility of our method with case studies.

### Consensus Monte Carlo for Random Subsets using Shared Anchors

♦Peter Mueller<sup>1</sup>, Yang Ni<sup>2</sup> and Yuan Ji<sup>3</sup>

<sup>1</sup>UT Austin

<sup>2</sup>TX A&M

<sup>3</sup>U. Chicago

pmueller@math.utexas.edu

We present a consensus Monte Carlo algorithm that scales existing Bayesian nonparametric models for clustering and feature allocation to big data. The algorithm is valid for any prior on random subsets such as partitions and latent feature allocation, under essentially any sampling model. Motivated by three case studies, we focus on clustering induced by a Dirichlet process mixture sampling model, inference under an Indian buffet process prior with a binomial sampling model, and with a categorical sampling model. We assess the proposed algorithm with simulation studies and show results for inference with three datasets: an MNIST image dataset, a dataset of pancreatic cancer mutations, and a large set of electronic health records (EHR).

### Connectivity Regression for heterogeneous networks

Jeffrey Morris

University of Pennsylvania

jeffrey.morris@penmedicine.upenn.edu

Abstract: Historically, much of statistics focuses on mean differences, while an important problem in modern science that has received less attention is the assessment of how associations across variables vary across subjects. One example is gene networks, whereby the subject-specific associations of molecules in important molecular pathways contain important biological information about the subjects, and another is in functional connectivity, whereby subject-specific associations of brain region activation may indicate important neurological mechanisms or dysfunctions. In this talk, we discuss new methods for performing regression analyses on association networks for high-dimensional multivariate data, in which we take data involving subject-specific networks and fit regression models and then assess through global tests which covariates affect the networks and local tests indicating which network edges vary significantly across covariates. The key to our novel modeling framework is projection to an alternative space where Gaussianity is justified, we can perform a parallel set of regression analyses on the real line while automatically constraining positive definiteness in the original space for all values of predictors, and we can gain efficiency by learning and accounting for second-order dependencies across network edges. We apply this approach to Human Connectome project data, and demonstrate the clear advantages of our method over existing Naïve approaches. This work introduces a new general regression framework to relate subject-specific graphs to discrete or continuous covariates.

### Session 48: Innovations in Statistical Machine Learning

#### Salient structure identification in complex networks by spectral periphery filtering

♦Tianxi Li<sup>1</sup>, Elizaveta Levina<sup>2</sup> and Ji Zhu<sup>2</sup>

<sup>1</sup>University of Virginia

<sup>2</sup>University of Michigan

tianxili@virginia.edu

Complex networks have been intensively studied in the past fifteen years. In practice, the salient network structure of interest, instead of being directly observed, is often hidden in a larger network in which most structures are not informative. The noise and bias introduced by this overwhelming yet non-informative data can obscure the salient structure and limit the effectiveness of many network analysis methods. Traditionally, researchers treat this scenario as a core-periphery structure, and algorithms are designed to extract the core. Unfortunately, most of these methods rely on restrictive assumptions on both the core and the periphery components that seriously undermine their usefulness. We propose a random network model for the non-informative structure of networks without imposing a specific form for the core. Specifically, we assume that the non-informative nodes are connected to other nodes in a purely random pattern, while the core structure can take any informative pattern. Moreover, we propose an algorithm of core extraction. The algorithm is computationally efficient and comes with a theoretical guarantee of accuracy. We evaluated the proposed model in extensive simulation studies and also use it to extract core structures in a few real-world networks for downstream analysis.

#### Brain regions identified as being associated with verbal reasoning through the use of imaging regression via internal variation

Long Feng<sup>1</sup>, ♦Xuan Bi<sup>2</sup> and Heping Zhang<sup>3</sup>

<sup>1</sup>City University of Hong Kong

<sup>2</sup>University of Minnesota

<sup>3</sup>Yale University

xbi@umn.edu

Brain-imaging data have been increasingly utilized to understand intellectual disabilities. Despite significant progress in biomedical research, the mechanisms for most of the intellectual disabilities remain unknown. Finding the underlying neurological mechanisms has proved difficult, especially in children due to the rapid development of their brains. We investigate verbal reasoning, which is a reliable measure of an individual's general intellectual abilities, and develop a class of high-order imaging regression models to identify brain subregions which might be associated with this specific intellectual ability. A key novelty of our method is to take advantage of spatial brain structures, and specifically the piecewise smooth nature of most imaging coefficients in the form of high-order tensors. Our approach provides an effective and urgently needed method for identifying brain subregions potentially underlying certain intellectual disabilities. The idea behind our approach is a carefully constructed concept called Internal Variation (IV). The IV employs tensor decomposition and provides a computationally feasible substitution for Total Variation (TV), which has been considered suitable to deal with similar problems but may not be scalable to high-order tensor regression. Before applying our method to analyze the real data, we conduct comprehensive simulation studies to demonstrate the validity of our method in imaging signal identification. Next, we present our results from the analysis of a dataset based on the Philadelphia Neurodevelopmental Cohort for which we pre-processed the data including re-orienting, bias-field correcting, extracting, normalizing and registering the magnetic resonance images

from 978 individuals. Our analysis identified a subregion across the cingulate cortex and the corpus callosum as being associated with individuals' verbal reasoning ability, which, to the best of our knowledge, is a novel region that has not been reported in the literature. This finding is useful in further investigation of functional mechanisms for verbal reasoning.

### Penalized likelihood estimation under distance-to-set penalties via majorization-minimization

Jason Xu

Duke University  
jason.q.xu@duke.edu

The majorization-minimization (MM) principle generalizes expectation-maximization (EM) algorithms to settings beyond missing data. Like EM, the idea relies on transferring optimization of a difficult objective (i.e. the likelihood under missing data) to a sequence of simpler subproblems (i.e. maximizing the expectation of the likelihood under complete data). We discuss MM approaches to regression problems under general constraints using distance-to-set penalties, making use of the recent proximal distance principle. Through this lens, we revisit sparse covariance estimation and high-dimensional regression under canonical constraints such as sparsity and rank restriction. We present strong empirical performance on several data examples and convergence guarantees even for non-convex objectives.

### Unsupervised Meets Supervised: Clustering of Regression Functions from Datasets

◆ Chenglong Ye<sup>1</sup> and Jie Ding<sup>2</sup>

<sup>1</sup>University of Kentucky

<sup>2</sup>University of Minnesota  
chenglong.ye@uky.edu

Motivated by the urgent need of processing massive information from distributed datasets, we present a method for clustering supervised relations between the response variable and predictors. In many data-processing scenarios of interest, the method enables reliable data integration, efficient-meta processing of data, and faster distributed computing. Our proposed solution allows each dataset to match a set of candidate parametric or nonparametric methods, and the clustering is applied to the selected methods. We prove theoretical guarantees, and discuss its applications to a variety of learning tasks including data integration, classical data-level clustering, and federated learning. Experimental studies using both synthetic and real data show remarkable performance (e.g. 38% reduction of mean squared error in the CT scan data) and significant computational advantage of the proposed method.

## Session 49: Advances in Clinical Trial Statistics

### Bayesian Optimal Phase II Design for Randomized Clinical Trials

◆ Yujie Zhao, Bo Yang, Jack J. Lee and Ying Yuan

The University of Texas MD Anderson Cancer Center  
yujie.zhao@uth.tmc.edu

Randomized clinical trials is gold standard to evaluate the efficacy of an experimental treatment. We propose a flexible Bayesian optimal phase II (BOP2) design for 2-arm randomized trials. The proposed 2-arm BOP2 design is flexible and can handle various types of endpoints, including binary, co-primary endpoint, and toxicity and efficacy endpoints under a unified framework. It also allows users to specify the number and timing of interim analyses to meet

the clinical needs. While enjoying the flexibility of Bayesian adaptive designs, the 2-arm BOP2 design explicitly controls type I error rate and is optimal in maximizing power, thereby ensuring desirable frequentist operating characteristics. Another important advantage of the 2-arm BOP2 design is that its decision rule can be tabulated and included in the trial protocol prior to the commence of the trial. To conduct the trial, no complicated Bayesian calculation is needed, and clinicians can simply look up the table and make go/no-go decisions. Simulation studies show that the 2-arm BOP2 design has desirable operating characteristics compared to group sequential design. Easy-to-use online application is freely available at [www.trialdesign.org](http://www.trialdesign.org) to facilitate the use of the BOP2 design in clinical trials.

### Elastic Meta-analytic-predictive Prior for Dynamically Borrowing Information from Historical Data with Application to Biosimilar Clinical Trials

◆ Wen Zhang<sup>1</sup>, Jean Pan<sup>2</sup> and Ying Yuan<sup>3</sup>

<sup>1</sup>The University of Texas Health Science Center at Houston

<sup>2</sup>Amgen, Inc.

<sup>3</sup>The University of Texas MD Anderson Cancer Center  
wen.zhang@uth.tmc.edu

A biosimilar is a biological product that is highly similar to and has no clinically meaningful differences from an approved reference product. Focusing on two-arm randomized clinical trials that aim to establish the equivalence between a test biosimilar product and the reference product, we propose the elastic meta-analytic-predictive (EMAP) prior method to leverage rich historical data available on the reference product to improve the power of the biosimilar trials. We first extract the prior information from multiple historical studies through meta-analysis, and then we discount the resulting meta-analytic-predictive (MAP) prior adaptively according to the congruence between the historical reference data and the trial reference arm data. We measure the congruence between the historical reference data and the trial reference arm data using the posterior predictive probability, and we achieve dynamic information borrowing by discounting the MAP prior using the elastic function of the congruence measure. The EMAP prior method encourages strong information borrowing when trial reference arm data are congruent to historical reference data, and forbids information borrowing when a substantial discrepancy exists between historical and current trial data. Extensive simulation studies show that the EMAP prior outperforms existing methods. The EMAP prior generates comparable or higher power and provides better-controlled type I errors. We illustrate the proposed methodology using two trial examples.

### Statistical Considerations in Clinical Trial Design with Event-free Survival as the Primary Efficacy Endpoint

◆ Yiming Zhang<sup>1</sup>, Tu Xu<sup>2</sup>, Meredith Goldwasser<sup>3</sup> and Vickie Zhang<sup>3</sup>

<sup>1</sup>University of Connecticut

<sup>2</sup>Vertex Pharmaceuticals Inc.

<sup>3</sup>Agios Pharmaceuticals Inc.  
yiming.3.zhang@uconn.edu

In late-phase confirmatory clinical trials in the oncology field, time-to-event (TTE) endpoints are commonly used as primary endpoints for establishing evidence on efficacy of investigational therapies. Among these TTE endpoints, overall survival (OS) is always considered the gold standard. However, OS data can take years to mature, and its measurement of efficacy can be confounded by the use of post-treatment rescue therapies or supportive care. Therefore, to accelerate the development process and better characterize the treatment effect of new investigational therapies, other TTE endpoints

such as progression-free survival and event-free survival (EFS) are applied as primary efficacy endpoints in some confirmatory trials, either as a surrogate for OS or as a direct measure of clinical benefit. For evaluating novel treatments for acute myeloid leukemia, EFS has been gradually recognized as a direct measure of clinical benefit. Nevertheless, the application of an EFS endpoint is still controversial mainly due to the debate surrounding definition of treatment failure (TF) events. In this study, we investigate the EFS endpoint with the most conservative definition on the timing of TF, which is Day 1 since randomization. Specifically, the corresponding non-proportional hazard pattern of the EFS endpoint makes the power of the log-rank test not necessary to be monotonically increasing with the event size. We explore the operating characteristics of the EFS endpoint using both analytical and numerical approaches with the goal to provide more insights on the trial design with EFS as the primary efficacy endpoint.

### **Trials of Targets**

♦ *Margret Erlendsdottir and Forrest Crawford*

Yale School of Public Health

margret.erlendsdottir@yale.edu

Randomized controlled trials (RCTs) are used to estimate the causal effect of a treatment on a health outcome of interest in a patient population. Often the specified treatment in an RCT is a medical intervention - such as a drug or procedure - experienced directly by the patient. Sometimes the "treatment" in an RCT is a target - such as a goal biomarker measurement - that the patient's physician attempts to reach using available medications or procedures. Large RCTs of targets are common in clinical research, and trials have been conducted to compare targets in the management of hypertension, diabetes, anemia, and acute respiratory distress syndrome. However, different RCTs intended to evaluate the same targets have produced conflicting recommendations and meta-analyses that aggregate results of trials of targets have been inconclusive. In this paper, we use principles of causal reasoning to explain why RCTs of targets conducted in different patient and physician populations can arrive at starkly different results, and why meta-analyses may yield invalid conclusions. We describe four key threats to the causal validity of trials of targets: 1) intention-to-treat analysis that conflates the effects of assignment to a biomarker target and medical treatments actually delivered to the patient; 2) incomparability in results across trials of targets, posing significant challenges for aggregating trials of targets using meta-analysis; 3) time-varying adaptive treatment strategies employed by physicians; and 4) Goodhart's law, "when a measure becomes a target, it ceases to be a good measure." We demonstrate the importance of comparing treatment strategies, in addition to targets, in randomized and observational studies, to generate useful and evidence-based clinical guidelines for physicians. We illustrate these findings using evidence from nine RCTs of blood pressure targets for management of hypertension.

### **Analysis of crossover designs with nonignorable dropout**

♦ *Xi Wang and Chinchilli Vernon*

Pennsylvania State University College of Medicine

xzw149@psu.edu

Clinical trials that invoke a crossover design usually require a higher level of participant adherence compared with parallel designs. Non-ignorable missing data are likely due to a long follow-up time and a high proportion of missing. However, very little work has been performed with respect to nonignorable missing data within the framework of crossover designs. This paper addresses the analysis of crossover designs with nonignorable dropout. We study

non-replicated crossover designs and replicated designs separately. With a primary objective of comparing the treatment mean effects, we jointly model the longitudinal measures and discrete time to dropout. We propose shared-parameter models and mixed-effects selection models. We adapt a linear-mixed effects model as the conditional model for the longitudinal outcomes. We invoke a discrete-time hazards model with a complementary log-log link function for the conditional distribution of time to dropout. We apply maximum likelihood for parameter estimation. We perform simulation studies to investigate the robustness of our proposed approaches under various missing data mechanisms. We then apply the approaches to two examples with a continuous outcome and one example with a binary outcome using existing software. In future work, we will implement a controlled multiple imputation method as a sensitivity analysis of the missing data assumption.

### **Semiparametric isotonic regression analysis for risk assessment under nested case-control and case-cohort designs**

♦ *Wen Li<sup>1</sup>, Ruosha Li<sup>2</sup>, Ziding Feng<sup>3</sup> and Jing Ning<sup>4</sup>*

<sup>1</sup>The University of Texas Health Science Center

<sup>2</sup>The University of Texas Health Science Center at Houston

<sup>3</sup>Fred Hutchinson Cancer Research Center

<sup>4</sup>The University of Texas MD Anderson Cancer Center

liwenmoi@gmail.com

Two-phase sampling designs, including nested case-control and case-cohort designs, are frequently utilized in large cohort studies involving expensive biomarkers. To analyze data from two-phase designs with a binary outcome, parametric models such as logistic regression are often adopted. However, when the model assumptions are not valid, parametric models may lead to biased estimation and risk evaluation. In this paper, we propose a robust semiparametric regression model for binary outcomes and an easy-to-implement computational procedure that combines the pool-adjacent violators algorithm with inverse probability weighting. The asymptotic properties are established, including consistency and the convergence rate. Simulation studies show that the proposed method performs well and is more robust than logistic regression methods. We demonstrate the application of the proposed method to real data from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial.

### **TITE-BOIN12: A Bayesian Adaptive Design to Find the Optimal Biological Dose with Late-onset Toxicity and Efficacy**

♦ *Yanhong Zhou, Ruitao Lin, Jack Lee and Ying Yuan*

The University of Texas MD Anderson Cancer Center

yanzhou03@gmail.com

In the era of immunotherapies and targeted therapies, the focus of early phase clinical trials has shifted from finding the maximum tolerated dose to the optimal biological dose (OBD) that maximizes the toxicity-efficacy trade-off. One major impediment to using adaptive designs to find the OBD is that efficacy or/and toxicity are often late-onset, disabling the decision rules of the designs to be applied in real time to treat new patients. To address this issue, we propose the TITE-BOIN12 design to find the OBD with late-onset toxicity and efficacy. As an extension of BOIN12 design, the TITE-BOIN12 design uses the utility to quantify the toxicity-efficacy trade-off. We considered two approaches, Bayesian data augmentation and approximated likelihood method, to enable real-time decision making when some patients' toxicity and efficacy outcomes are still pending. Extensive simulations show that compared to some existing designs, TITE-BOIN12 significantly shortens the trial duration, with comparable or higher accuracy to identify the OBD and a

lower risk of overdosing patients. To facilitate the use of the TITE-BOIN12 design, we develop user-friendly software freely available at [www.trialdesign.org](http://www.trialdesign.org).

## Session 50: Bayesian Statistics

### Bayesian Modeling of Spatial Transcriptomics Data via a Modified Potts Model

♦ *Xi Jiang*<sup>1</sup>, *Qiwei Li*<sup>2</sup> and *Guanghua Xiao*<sup>3</sup>

<sup>1</sup>Southern Methodist University

<sup>2</sup>The University of Texas at Dallas

<sup>3</sup>The University of Texas Southwestern Medical Center  
[xij@smu.edu](mailto:xij@smu.edu)

As recently developed techniques, spatial molecular profiling (SMP) provides opportunities to understand characterization of cells and their spatial organizations comprehensively. Two main approaches of SMP are image-based and sequencing-based techniques. Our study focused on analyzing sequencing-based SMP data and aimed at identifying genes that display spatial expression patterns given their expression count data and location information. We introduced a Bayesian modelling of spatial transcriptomics data via a modified Potts model by using Hamiltonian energy measurement to define spatial pattern among gene expression levels. For each gene, expression counts data are transformed to the states of under and over-expression levels on a lattice. Then, the interaction parameter in a Potts model with external field is estimated via Bayesian modeling and whether the gene has spatial expression pattern can be inferred. Our simulation results showed high power while remaining relatively low Type I error. We applied our model on mouse olfactory bulb data and a new pattern of gene expression can be detected compared with other existing methods, such as SPARK and SpatialDE. Since our method projects count data into binary gene-expression levels and prevents the selection of ad hoc kernels, it performs a more robust inference compared with other methods.

### A Bayesian Nonparametric Approach for Inferring Drug Combination Effects on Mental Health in People with HIV

♦ *Wei Jin*<sup>1</sup>, *Yang Ni*<sup>2</sup>, *Leah Rubin*<sup>3</sup>, *Amanda Spence*<sup>4</sup> and *Yanxun Xu*<sup>1</sup>

<sup>1</sup>Johns Hopkins University

<sup>2</sup>Texas A&M University

<sup>3</sup>Johns Hopkins University School of Medicine

<sup>4</sup>Georgetown University  
[wjin@jhu.edu](mailto:wjin@jhu.edu)

Although combination antiretroviral therapy (ART) is highly effective in suppressing viral load for people with HIV (PWH), many ART agents may exacerbate central nervous system (CNS)-related adverse effects including depression. Therefore, understanding the effects of ART drugs on the CNS function, especially mental health, can help clinicians personalize medicine with less adverse effects for PWH and prevent them from discontinuing their ART to avoid undesirable health outcomes and increased likelihood of HIV transmission. The emergence of electronic health records offers researchers unprecedented access to HIV data including individuals' mental health records, drug prescriptions, and clinical information over time. However, modeling such data is very challenging due to high-dimensionality of the drug combination space, the individual heterogeneity, and sparseness of the observed drug combinations. We develop a Bayesian nonparametric approach to learn drug combination effect on mental health in PWH adjusting for

socio-demographic, behavioral, and clinical factors. The proposed method is built upon the subset-tree kernel method that represents drug combinations in a way that synthesizes known regimen structure into a single mathematical representation. It also utilizes a distance-dependent Chinese restaurant process to cluster heterogeneous population while taking into account individuals' treatment histories. We evaluate the proposed approach through simulation studies, and apply the method to a dataset from the Women's Interagency HIV Study, yielding interpretable and promising results. Our method has clinical utility in guiding clinicians to prescribe more informed and effective personalized treatment based on individuals' treatment histories and clinical characteristics.

### Informative sampling of Bayesian inference for continuous repeated measurement response

*Helen Engle*

UTH

[helen.engle@uth.tmc.edu](mailto:helen.engle@uth.tmc.edu)

Survey data is usually obtained under complex sampling design. Analyzing the data as if it came from a simple random design would yield biased results. To take into account the sampling design each survey data point is released with a sampling weight proportional to the number of individuals in the population that the data point represents and that is also adjusted for non-response. Sometimes the sampling weights are correlated with the response variable of interest. If this is the case we say that the sampling design is informative (with respect to the response). For instance, individuals would be more likely to respond to the survey if they have a positive feedback. That is to say, the survey sample we are analyzing is not a simple random sample. Leon-Novelo and Savitsky (2019) introduce a Bayesian model-base[1] inference that takes into account sampling weights, adjusting for the sampling design. Leon-Novelo and Savitsky focus on continuous response, here we are extending their approach to repeated measure of continuous response. We compare the performance of our approach with competing methods via simulation. The comparison is carried out in terms of bias, mean square error (MSE), coverage and length of credible intervals. Finally, we demonstrate the methods by applying them to National Health and Nutrition Examination Survey (NHANES) dataset.

### Evaluating Short-term Forecast among Different Epidemiological Models under a Bayesian Framework

*Qiwei Li*<sup>1</sup>, ♦ *Tejasv Bedi*<sup>1</sup>, *Guanghua Xiao*<sup>2</sup> and *Yang Xie*<sup>3</sup>

<sup>1</sup>University of Texas at Dallas

<sup>2</sup>UT Southwestern Medical Center

<sup>3</sup>UT Southwestern Medical Center.  
[txb180007@utdallas.edu](mailto:txb180007@utdallas.edu)

Forecasting of COVID-19 daily confirmed cases has been one of the several challenges posed on the governments and health sectors on a global scale. To facilitate informed public health decisions, the concerned parties rely on short-term daily projections generated via predictive modeling. We calibrate stochastic variants of growth models and the standard SIR model into one Bayesian framework to evaluate their short-term forecasts. In summary, it was noted that none of the models proved to be golden standards across all the regions in their entirety, while all outperformed ARIMA in a predictive capacity as well as in terms of interpretability.

### Bayesian Landmark-Based Shape Analysis of Tumor Pathology Images

♦ *Cong Zhang*, *Kelli Palmer*, *Min Chen*, *Michael Zhang* and *Qiwei Li*

University of Texas at Dallas

cxz163430@utdallas.edu

Lung cancer has been ranked as the leading cause of death from cancer around the world. Accurate diagnosis and prognosis prediction are critical for providing guidance for the clinical therapy strategies. Pathology images has become a routine clinical procedure and recognized as the gold standard. Traditional manual inspection by pathologist experts is becoming impractical with the rapid increasing sample size of pathology images. Automatic tumor shape analysis of pathology images with both accuracy and efficiency is in need. Here we proposed a Bayesian LAndmark-based Shape Analysis (BayesLASA) framework for tumor pathology images. We use polygonal chain to represent tumor shape data and MCMC algorithms to sample from the posterior distribution. We demonstrate the improved accuracy and time scalability of our automatic landmark detection model in simulated datasets by comparisons with previous approaches. We further proposed a "skeleton"-referenced tumor shape boundary characterizing analysis method and applied it in a case study of 246 pathology images from 143 non-small cell lung cancer patients. Two sets of features were put forward, one based on direct surface roughness measurement and one based on Hidden Markov model. The case study shows that both sets of features under the "skeleton" paradigm were able to capture the heterogeneity property of tumor shape boundary and demonstrated their predictive roles of prognosis.

#### **BayesSMILES: Bayesian Segmentation ModelIng for Longitudinal Epidemiological Studies**

♦ *Shuang Jiang<sup>1</sup>, Quan Zhou<sup>2</sup>, Xiaowei Zhan<sup>3</sup> and Qiwei Li<sup>4</sup>*

<sup>1</sup>Southern Methodist University

<sup>2</sup>Texas A&M University

<sup>3</sup>University of Texas Southwestern Medical Center

<sup>4</sup>The University of Texas at Dallas  
shuangj@smu.edu

Coronavirus disease 2019 (COVID-19) is a pandemic. To characterize the disease transmissibility, we propose a Bayesian change point detection model using daily actively infectious cases. Our model is built upon a Bayesian Poisson segmented regression model that can 1) capture the epidemiological dynamics under the changing conditions caused by external or internal factors; 2) provide uncertainty estimates of both the number and locations of change points; 3) adjust any explanatory time-varying covariates. Our model can be used to evaluate public health interventions, identify latent events associated with spreading rates, and yield better short-term forecasts.

#### **Bayesian Functional Regression on Manifold With Application to Infant Cortical Thickness**

♦ *Ye Emma Zohner<sup>1</sup> and Jeffrey Morris<sup>2</sup>*

<sup>1</sup>Rice University

<sup>2</sup>University of Pennsylvania  
emma.zohner@rice.edu

Biomedical research increasingly involves the collection of high-dimensional complex data. Often the data are functional data where observations are a set of curves that are sampled on a fine grid. Moreover, the curves may be defined on a manifold. We propose a framework that studies various bases to determine the best representation for complex functional data where the domain is a manifold, and we model the data via Bayesian functional mixed model. This methodology is motivated by infant cortical thickness data, but the modeling strategy we present can be applied to any functional mixed modeling when the object lives on a manifold.

#### **Bayesian and Unsupervised Machine Learning Machines for Jazz Music Analysis**

*Qiuyi Wu*

University of Rochester  
qiuyi.wu@urmc.rochester.edu

Extensive studies have been conducted on both musical scores and audio tracks of western classical music with the finality of learning and detecting the key in which a particular piece of music was played. Both the Bayesian Approach and modern unsupervised learning via latent Dirichlet allocation have been used for such learning tasks. In this research work, we venture out of the western classical genre and embrace and explore jazz music. We consider the musical score sheets and audio tracks of some of the giants of jazz like Duke Ellington, Miles Davis, John Coltrane, Dizzy Gillespie, Wes Montgomery, Charlie Parker, Sonny Rollins, Louis Armstrong, Gil Evans, Bill Evans, Dave Brubeck, Thelonious Monk. We specifically employ Bayesian techniques and modern topic modeling methods and a combination of both to explore tasks such as: automatic improvisation detection, genre identification, key learning (how many keys do the giants of jazz tended to play in, and what are those keys) and even elements of the mood of the piece.

#### **Double spike Dirichlet priors for structured weighting**

♦ *Huiming Lin and Meng Li*

Rice University  
hl68@rice.edu

Assigning weights to a large pool of objects is a fundamental task in a wide variety of applications. In this article, we introduce a concept of structured high-dimensional probability simplexes, whose most components are zero or near zero and the remaining ones are close to each other. Such structure is well motivated by 1) high-dimensional weights that are common in modern applications, and 2) ubiquitous examples in which equal weights—despite their simplicity—often achieve favorable or even state-of-the-art predictive performances. This particular structure, however, presents unique challenges both computationally and statistically. To address these challenges, we propose a new class of double spike Dirichlet priors to shrink a probability simplex to one with the desired structure. When applied to ensemble learning, such priors lead to a Bayesian method for structured high-dimensional ensembles that is useful for forecast combination and improving random forests, while enabling uncertainty quantification. We design efficient Markov chain Monte Carlo algorithms for easy implementation. Posterior contraction rates are established to provide theoretical support. We demonstrate the wide applicability and competitive performance of the proposed methods through simulations and two real data applications using the European Central Bank Survey of Professional Forecasters dataset and a UCI dataset.

#### **Session 51: Recent developments in AI and its Applications**

##### **Multi-scale affinities with missing data**

♦ *Min Zhang<sup>1</sup>, Gal Mishne<sup>2</sup> and Eric Chi<sup>1</sup>*

<sup>1</sup>North Carolina State University

<sup>2</sup>University of California San Diego  
mzhang27@ncsu.edu

In many machine learning problems, we need to compute weights that capture the proximity information of the data matrix. The choice of weights can dramatically affect the effectiveness of the algorithm. Nonetheless, the problem of choosing weights is not

given enough study, and weights are usually picked by heuristics. If the complete matrix is observed, employing the Gaussian kernel affinities is a common practice to quantify the local similarity between pairs of rows and pairs of columns. However, in the presence of missing data, computing the weights is more difficult and complicated. In this paper, we propose a new method to construct row and column affinities even when data is missing by building off the co-clustering technique. This metric takes advantage of solving the optimization problem for multiple pairs of cost parameters and filling in the missing values with increasingly smooth estimates. It exploits the coupled similarity structure among both the rows and columns of a data matrix. We show these affinities can be used to perform tasks such as data imputation, matrix completion on graphs, and clustering.

### **GPU accelerated statistical methods through a deep learning framework**

*Shikun Wang*

The University of Texas MD Anderson Cancer Center  
shikunw666@gmail.com

Author: Shikun Wang, Zhao Li, Lan Lan, Wenjin Zheng, Liang Li  
Abstract: In longitudinal cohort studies, it is often of interest to predict the risk of a terminal clinical event using accumulating longitudinal predictor data among those patients who are still at-risk for the terminal event. The at-risk patient population may change over time, and so is the association between predictors and the outcome. This dynamic prediction problem has received increasing interest in the literature, but there remain computational challenges in its analysis. The widely used joint model of longitudinal and survival data often suffers intensive computation or excessive model fitting time, due to numerical optimization and the analytically intractable high-dimensional integral in the likelihood function. This problem is exacerbated when the model is fit to a large dataset or the model involves multiple longitudinal predictors with nonlinear trajectories. In this paper, we address this problem from an algorithmic perspective, by proposing a novel two-stage estimation procedure, and from a computing perspective, by using Graphics Processing Unit (GPU) programming. The latter is implemented through PyTorch, an emerging deep learning framework. Our numerical studies demonstrate that our proposed algorithm and software can substantially speed up the estimation of the joint model, particularly with large datasets. We also found that accounting for the nonlinearity in longitudinal predictor trajectories can improve the prediction accuracy in comparison to joint models that ignore nonlinearity. An open-source software for model development and dynamic prediction is available. Keywords: Dynamic prediction, electronic health records, Graphics Processing Unit (GPU) computing, joint modeling, longitudinal and survival data, numerical integration, parallel computing

### **Parameter Estimation and Inference of Spatial Autoregressive Model by Stochastic Gradient Descent**

♦ *Gan Luan and Ji Meng Loh*

New Jersey Institute of Technology  
gl238@njit.edu

Many data contain spatial components, and it is important to consider spatial correlation in modeling and parameter estimation. Spatial autoregressive (SAR) model is often used to modeling these data. Parameters of SAR model are mainly estimated by maximum likelihood method (based on profile likelihood). However, no closed form of MLE exists, and it cannot scale up well due to heavy computation involved in numerical methods used for parameter estimation.

Stochastic gradient descent (SGD) is a desirable method for model parameter estimation in large-scale data and online learning settings since it goes through the data in only one pass. Although many studies regarding SGD have been conducted, application of SGD for spatial models is still not common. In this project, we consider spatial lattice data and use averaged SGD for model parameter estimation. Data in SAR model are correlated, so a new SGD algorithm is proposed to accommodate this data dependence. Also, a bootstrap procedure is used to conduct inferences based on SGD estimator. This inference procedure updates SGD estimates, and at the same time generates many randomly perturbed SGD estimates for each observation. These perturbed estimates can be used to produce confidence intervals. We used simulations to study the performance of this parameter estimation and confidence interval construction algorithm. Simulation results suggest that SGD estimates converge to true values, however, confidence interval coverage is not close to the desired level for the spatial parameter. We proposed two methods for improving the coverage of confidence intervals. Last we studied the asymptotic properties of SGD estimates.

### **Deep Learning for Quantile Regression: DeepQuantreg**

♦ *Yichen Jia and Jong-Hyeon Jeong*

University of Pittsburgh  
yij22@pitt.edu

The computational prediction algorithm of neural network, or deep learning, has drawn much attention recently in statistics as well as in image recognition and natural language processing. Particularly in statistical application for censored survival data, the loss function used for optimization has been mainly based on the partial likelihood from Cox's model and its variations to utilize existing neural network library such as Keras, which was built upon the open source library of TensorFlow. This paper presents a novel application of the neural network to the quantile regression for survival data with right censoring, which is adjusted by the inverse of the estimated censoring distribution in the check function. The main purpose of this work is to show that the deep learning method could be flexible enough to predict nonlinear patterns more accurately compared to the traditional method even in low-dimensional data, emphasizing on practicality of the method for censored survival data. Simulation studies were performed to generate nonlinear censored survival data and compare the deep learning method with the traditional quantile regression method in terms of prediction accuracy. The proposed method is illustrated with a publicly available breast cancer data set with gene signatures.

### **Algorithmic Regularized Fusion Minimization-Majorization Method for Clustering Histogram Data**

♦ *Xu Han and Eric Chi*

North Carolina State University  
xhan22@ncsu.edu

Convex clustering is a popular way to do clustering, but with a traditional structure based on the Euclidean distance, it may wrongly specify clusters for histogram data where perturbations or shifts always happen. Considering the stability of Earth Mover's Distance (EMD) in measuring histogram data, we replace the Euclidean distance with EMD in the penalty of a convex clustering problem. To avoid the computational cost of multiple EMDs, we define another distance to approximate EMD by cutting histogram into different pieces. We propose a Minimization-Majorization algorithm with a heuristic fusion setting and the algorithmic regularized path to improve computational efficiency.

### **Appearance-free Tripartite Matching for Multiple Object**

**Tracking**

♦ *Lijun Wang*<sup>1</sup>, *Yanting Zhu*<sup>2</sup>, *Jue Shi*<sup>2</sup> and *Xiaodan Fan*<sup>3</sup>

<sup>1</sup>Chinese University of Hong Kong

<sup>2</sup>Hong Kong Baptist University

<sup>3</sup>The Chinese University of Hong Kong

ljwang@link.cuhk.edu.hk

Multiple Object Tracking (MOT) detects the trajectories of multiple objects given an input video, and it has become more and more popular in various research and industry areas, such as cell tracking for biomedical research and human tracking in video surveillance. We target at the general MOT problem regardless of the object appearance. The appearance-free tripartite matching is proposed to avoid the irregular velocity problem of traditional bipartite matching. The tripartite matching is formulated as maximizing the likelihood of the state vectors constituted of the position and velocity of objects, and a dynamic programming algorithm is employed to solve such maximum likelihood estimate (MLE). To overcome the high computational cost induced by the vast search space of dynamic programming, we decompose the space by the number of disappearing objects and propose a reduced-space approach by truncating the decomposition. Extensive simulations have shown the superiority and efficiency of our proposed method. We also applied our method to track the motion of natural killer cells around tumor cells in a cancer research.

**Recommender system of scholarly papers using public datasets**

♦ *Jie Zhu*<sup>1</sup>, *Braja Patra*<sup>2</sup>, *Hulin Wu*<sup>1</sup> and *Ashraf Yaseen*<sup>1</sup>

<sup>1</sup>The University of Texas Health Science Center at Houston

<sup>2</sup>The University of Texas Health Science Center at Houston; Weill Cornell Medicine

jie.zhu@uth.tmc.edu

The exponential growth of public datasets in the era of Big Data demands new solutions for making these resources findable and reusable. Therefore, a scholarly recommender system for public datasets is an important tool in the field of information filtering. It will aid scholars in identifying prior and related literature to datasets, saving their time, as well as enhance the datasets reusability. In this work, we developed a scholarly recommendation system that recommends research-papers, from PubMed, relevant to public datasets, from Gene Expression Omnibus (GEO). Different techniques for representing textual data are employed and compared in this work. Our results showed that term-frequency based methods (BM25 and TF-IDF) outperformed all others including popular Natural Language Processing embedding models such as doc2vec, ELMo, and BERT.

**Extracting Clinically Meaningful Features for the Analysis of Tumor Pathology Images**

*Esteban Fernandezmorales*

The University of Texas at Dallas

esteban.fernandezmorales@utdallas.edu

With the advance of imaging technology, digital pathology imaging of tumor tissue slides is becoming a routine clinical procedure for cancer diagnosis. This process produces massive imaging data that capture histological details in high resolution. Recent developments in deep-learning methods have enabled us to automatically detect and characterize the tumor regions in pathology images on a large scale. From each identified tumor region, we extracted 40 well-defined descriptors that quantify its shape, geometry, and topology. We demonstrated the association between those descriptor features and patient survival outcome in lung adenocarcinoma patients from the National Lung Screening Trial (n=143). Besides, a descriptor-based prognostic model was developed and validated in an indepen-

dent patient cohort (n=321). This study proposes new insights into the relationship between tumor shape, geometrical, and topological features and patient prognosis.

**Session 52: Statistics in Genetics****Efficient odds ratio estimation using partial data audits in error-prone, observational HIV cohort data**

♦ *Sarah C. Lotspeich*<sup>1</sup>, *Bryan E. Shepherd*<sup>2</sup>, *Gustavo G. C. Amorim*<sup>1</sup>, *Pamela A. Shaw*<sup>3</sup> and *Ran Tao*<sup>1</sup>

<sup>1</sup>Vanderbilt University

<sup>2</sup>Vanderbilt University

<sup>3</sup>University of Pennsylvania

sarah.c.lotspeich@vanderbilt.edu

Persons living with HIV engage in clinical care often, so observational HIV research cohorts generate especially large amounts of routine clinical data. Increasingly, these data are being used in biomedical research, but available information can be error prone and biased statistical estimates can mislead results. The Caribbean, Central, and South America network for HIV epidemiology is one such cohort; fortunately, data audits have been conducted. Risk of AIDS defining event after initiating antiretroviral therapy is of clinical interest, expected to be associated with CD4 lab value and AIDS status. Error-prone values for 5109 patients were in the research database, and validated data were available (substantiated by clinical source documents) on only 117 patients. Instead of naive (unaudited) or complete case (audited) analysis, we propose a novel semi-parametric likelihood method using all available information (unaudited and audited) to obtain unbiased, efficient odds ratios with error prone outcome and covariates. Point estimates were farther from the null than the naive analysis, directionality agreed with the complete case analysis, but had narrower confidence intervals.

**Age-related alterations in fractal behaviors of respiratory signals**

♦ *Teng Zhang*<sup>1</sup>, *Xinzheng Dong*<sup>2</sup>, *Chen Chang*<sup>1</sup> and *Xiaohua Douglas Zhang*<sup>1</sup>

<sup>1</sup>University of Macau

<sup>2</sup>University of Macau, South China University of Technology

yb77605@um.edu.mo

In recent years, the fast development of medical devices makes the monitoring on respiratory signals more accurate and convenient. A large amount of continuous monitoring of respiratory signals is available for people to conduct further analysis. The fractal is a primary nonlinear characteristic of physiological data and some fractal analysis methods have been applied in physiological signals. Previous studies have found that the fractal behaviors exist in inter breath intervals (IBI) of respiratory signals. To explore the age-related alterations in fractal behaviors of respiratory signals, we compared the fractal behaviors of IBI in young and elderly. The data was downloaded from Fantasia database in PhysioNet which contains the respiratory signals of 10 young (21-34 years old) and 10 elderly (68-81 years old). We calculated the IBI series of the respiratory signals and analyzed the IBI series using Power Spectral Density (PSD) and Multifractal Detrended Fluctuation Analysis (MFDFA). Results show that the scaling exponents of PSD are significantly different between young and elderly and the Hurst exponents of MFDFA in young are higher than that in elderly. The age-related difference of fractal behaviors in respiratory signals may reflect the change of physiological self-regulation ability, and fractal has the potential to be a feature to detect some diseases.



### Something out of Nothing? The Influence of 0-0 Studies in Drug Safety Analysis

♦Zhaohu Fan<sup>1</sup>, Dungan Liu<sup>1</sup>, Yuejie Chen<sup>2</sup> and Nanhua Zhang<sup>1</sup>

<sup>1</sup>University of Cincinnati

<sup>2</sup>North Carolina State University  
fanzh@mail.uc.edu

Examining the safety of new drugs is important in clinical trials. In clinical studies, the occurrence of adverse events (e.g. deaths) is often rare. However, in the case of rare events, a single study may not produce sufficient power to detect meaningful signals. The meta-analysis, which synthesizes multiple studies, is widely used to analyze rare events data. When events are rare, some of the studies may observe zero events in both treatment and control groups. These studies are referred to as double-zero studies. The influence of double-zero studies has been researched in the literature, but it still remains unsettled. Some argued that in theory, these studies contain information for inference, whereas others do not observe their influence in numerical studies. This paper examines when and how double-zero studies contribute to inference in Bayesian analysis. Through extensive numerical studies, we demonstrate that 1) type I error can be significantly inflated, 2) the testing power can be significantly decreased, and 3) bias can be increased if double zero event studies are excluded from the analysis.

### Epidemiology characteristics of influenza A and B in Macau

Hoiman Ng<sup>1</sup>, Teng Zhang<sup>2</sup>, ♦Guoliang Wang<sup>2</sup>, Simeng Kan<sup>1</sup>, Guoyi Ma<sup>2</sup>, Zhe Li<sup>2</sup>, Chang Chen, Dandan Wang<sup>2</sup>, Mengin Wong<sup>1</sup> and Chiohang Wong<sup>1</sup>

<sup>1</sup>Kiang Wu Hospital

<sup>2</sup>University of Macau  
yb87623@umac.mo

Influenza caused by Influenza viruses is one of the major respiratory diseases in humans. Influenza A and B viruses are the main influenza virus that are widespread and cause epidemic in human. Influenza epidemics typically occur during the winter months in temperate regions, whereas influenza seasonality is very diverse in subtropical and tropical regions. Located in the tropics, Macau is a world-famous tourist city and has a large population mobility. Complex environmental and human factors affect the influenza epidemic in Macau. We collected influenza data in Macau from 2010 to 2018. Chi-square test and binary multivariable logistic regression were used to investigate the epidemiological characteristics of influenza A and B in Macau. A total of 104,874 samples with influenza-like illness (ILI) were collected. The samples consisted of 17973 cases (17.14%) of influenza A and 7274 cases (6.94%) of influenza B. We find that the influenza A occurred year-round, with the highest positive rate in winter. Influenza B showed seasonality, with the peak occurring in March and April of the spring season. We also found that Macau experienced three influenza pandemics from 2010 to 2018. The pandemic in July 2017 was caused by influenza A, while two pandemics in January-February 2012 and January-February 2018 were caused by influenza B. Finally, influenza infection was not related to gender and the odds of influenza was low in children (0-4 years old) and the elderly (>64 years old). Our study on the epidemiological characteristics of influenza in Macau can be helpful for the prevention of influenza and provide guidance on vaccine strain selection and vaccination time adjustment.

### Prevalence of Allergen Sensitization in Patients with Allergic Diseases in mainland China: A Four-year Retrospective Study

♦Dandan Wang<sup>1</sup>, Wenting Luo<sup>2</sup>, Teng Zhang<sup>1</sup>, Peiyan Zheng<sup>2</sup>, Dongliang Leng<sup>1</sup>, Baoqing Sun<sup>2</sup> and Xiaohua Zhang

<sup>1</sup>University of Macau

<sup>2</sup>National Clinical Research Center of Respiratory Disease  
yb97620@um.edu.mo

The population with allergic diseases increases rapidly in recent decades. Allergy not only leads to poor quality life in patients but also causes significant economic burden to the society. Study on prevalence of allergens contributes to the treatment and prevention of allergic diseases. Although many studies on the prevalence of allergens in China have been published, there is still lack of researches on the geographic distribution of allergens in mainland China. In this paper, we analyzed the prevalence patterns of serum allergen-specific IgE (sIgE) sensitization to 4 most common food allergens and 5 aeroallergens among 44156 patients with allergic diseases in seven geographic regions of mainland China from 2015 to 2018. Results show that prevalence of allergies vary in different regions, age groups, gender and seasons. Our findings may help clinical staffs to apply effective individualized treatment into unique patient group and direct researchers to conduct deeper studies on epidemiology of allergic diseases.

### Spatiotemporal profiling of COVID-19 epidemic in Hubei

Kuan Cheok Lei

University of Macau  
johnnylei@um.edu.mo

Coronavirus disease 2019 (COVID-19, previously known as novel coronavirus pneumonia, Wuhan pneumonia or Wuhan acute respiratory syndrome) is an epidemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, previously named as 2019-nCoV). Several clustered pneumonia cases were first reported in late December in Wuhan, Hubei province, China, but not until the mid of January of the following year when the qualitative assay for SARS-CoV-2 was out, and continuous increase in positives in the growing number of pneumonia patients in Wuhan, that SARS-CoV-2 was finally recognized as a pathogen acquirable by person-to-person transmission. As the outbreak occurs near the Chinese New Year travel season, a large-scale isolation policy was executed to prevent further spread of the virus, which resulted in the lockdown of Wuhan and other provincial regions of Hubei province in late January. Until 14 February, the number of confirmed cases reaches 64477, in which 62862 are confirmed in mainland China, in which 51986 (80.6% of total confirmed cases) are in Hubei province. In this study, we obtain epidemiological and demographical statistics from Health Commission of Hubei Province and other sources to analyze the spatiotemporal transmission of the epidemic in Hubei, and estimate the factors causing the differences of the confirmation numbers between Hubei and other provinces.

### Covariate adjustment in continuous biomarker assessment

♦Ziyi Li<sup>1</sup>, Zhenxing Guo<sup>1</sup>, Ying Cheng<sup>2</sup>, Peng Jin<sup>1</sup> and Hao Wu<sup>1</sup>

<sup>1</sup>Emory University

<sup>2</sup>Yunnan University  
ziyi.li@emory.edu

Continuous biomarkers are common for disease screening and diagnosis. To reach a dichotomous clinical decision, a threshold would be imposed to distinguish subjects with disease from non-diseased individuals. Among various performance metrics for a continuous biomarker, specificity at a controlled sensitivity level (or vice versa) is often desirable for clinical utility since it directly targets where the clinical test is intended to operate. Covariates, such as age, race, and sample collection, could impact the controlled sensitivity level in subpopulations and may also confound the association between biomarker and disease status. Therefore, covariate adjustment is important in such biomarker evaluation. In this paper, we

suggest to adopt a parsimonious quantile regression model for the diseased population, locally at the controlled sensitivity level, and assess specificity with covariate-specific control of the sensitivity. Variance estimates are obtained from a sample-based approach and bootstrap. Furthermore, our proposed local model extends readily to a global one for covariate adjustment for the receiver operating characteristic (ROC) curve over the sensitivity continuum. We demonstrate computational efficiency of this proposed method and restore the inherent monotonicity in the estimated covariate-adjusted ROC curve. The asymptotic properties of the proposed estimators are established. Simulation studies show favorable performance of the proposal. Finally, we illustrate our method in biomarker evaluation for aggressive prostate cancer.

### Session 53: Recent statistical advances in longitudinal and survival analysis

#### Subgroup Analysis of Longitudinal Profiles for Compositional Count Data

◆ *Chenyang Duan and Yuan Jiang*

Oregon State University  
duanc@oregonstate.edu

Competition-colonization trade-off is a stabilizing mechanism underlying species diversity in biological systems, especially those that are not in equilibrium. To test this mechanism, modern ecological studies collect longitudinal and compositional counts of the DNA sequences of taxonomically diagnostic genetic markers to measure the abundance of different species. A fundamental question in competition-colonization trade-off is to group the species according to the similarity of their biological roles in this mechanism, which can be characterized by the similarity of the longitudinal trajectories of their abundances. In this paper, we propose a novel method named COMPARING for subgroup analysis of longitudinal profiles for the species abundances in a biological system. In this method, Dirichlet-multinomial regression is used to link the compositional counts of the species to a longitudinal covariate revealing different longitudinal profiles for different species and subgroup analysis is performed to cluster similar longitudinal profiles together. We use the nonparametric B-spline smoothing techniques to estimate the longitudinal patterns and rely on a pairwise-distance penalization to identify subgroups with similar longitudinal patterns. We develop and implement the linearized alternating direction method of multipliers (L-ADMM) algorithm to estimate the parameters in the proposed model. Simulation studies demonstrate the advantage of COMPARING over its competitors in terms of the accuracy of recovering the underlying clusters of longitudinal trajectories. In addition, we apply COMPARING to a real dataset to reveal the co-existence of blood-borne parasites in African buffalo and demonstrate how the method successfully detects biologically meaningful subgroups of parasites for the competition-colonization trade-off.

#### Comparative Analysis of Haplotype Assembly Algorithms

◆ *Shuying Sun<sup>1</sup>, Flora Cheng<sup>2</sup>, Daphne Han<sup>3</sup>, Sarah Wei<sup>4</sup> and Alice Zhong<sup>5</sup>*

<sup>1</sup>Texas State University

<sup>2</sup>Westwood High School

<sup>3</sup>Kingwood High School

<sup>4</sup>Massachusetts Institute of Technology

<sup>5</sup>Clements High School  
ssun5211@yahoo.com

A haplotype is a set of DNA variants (e.g., the alleles of single nucleotide polymorphism) inherited together from one parent or one chromosome. Haplotype information is useful for studying genetic variation and can be used for disease association studies. Haplotype assembly means inferring haplotypes using DNA sequencing reads. Currently, there are a great number of haplotype assembly algorithms, and each has its own strengths and weaknesses. In this project, we focus on comparing several haplotype assembly algorithms: HapCUT2, PEATH, MixSIH, WhatsHap, MATChap, and SDhap. We use the NA12878 dataset to compare their accuracy and efficiency. To assess their efficiency, we compare the running time of each algorithm. To assess their accuracy, we perform pairwise comparisons between each pair of the six packages using R and Perl to generate haplotype block and single nucleotide variant (SNV) disagreement values. We also compare them using a switch error (switch distance), which is the number of positions where two chromosomes of a certain phase must be switched to match with the “true phase”. We find that HapCUT2, PEATH, MixSIH, and MATChap generate output files with similar numbers of SNVs and have relatively similar performance. HapCUT2 is very efficient, and its output has a high agreement rate with PEATH. In addition, we find that WhatsHap generates the output with a much larger number of SNVs. This large difference of SNV numbers causes WhatsHap to have high disagreement rates with other algorithms. We find that SDhap is less accurate and much slower than the other algorithms. All these findings provide new perspectives on the performance of currently available haplotype assembly algorithms and useful input for users of haplotype assembly algorithms.

#### Transcriptome analysis reveals lncRNA-mediated complex regulatory network response to DNA damage in the liver tissue of *Rattus norvegicus*

◆ *Chen Huang, Dong Liang Leng, Kuan Cheok Lei, Shi Xue Sun and Xiaohua Douglas Zhang*

University of Macau  
yb97648@um.edu.mo

DNA is prone to damages, which would result in genetic disorders and enhance risk of tumorigenesis. Hence, understanding the molecular mechanisms of DNA damage and repair will provide deep insights into tumorigenesis, carcinogenesis as well as the corresponding treatments. Aiming at investigating potential lncRNAs response against DNA damage, we performed a comprehensive transcriptomic analysis based on RNA sequencing data of the liver tissue from *Rattus norvegicus*, in which DNA damage was induced using aflatoxin B1, ifosfamide and N-nitrosodimethylamine. Through our analyses, numerous novel lncRNAs are identified for the first time, and differential network analysis discloses lncRNA-mediated regulatory networks related to DNA damage response. The result shows that these DNA damage-inducing chemicals might disrupt many lncRNA-mediated interactions involved in diverse biological processes and pathways, e.g., immune function and cell adhesion. In contrast, the host might also activate a few RNA interactions in response to DNA damage, involving response to drug and regulation of cell cycle.

#### Development of novel and robust method for analyzing single-cell RNA sequencing data

◆ *Min Deng and Xiao Hua Douglas Zhang*

Faculty of Health Sciences  
yb87624@um.edu.mo

Abstract The advent of single-cell RNA sequencing (scRNA-seq) technologies has made it possible to study the transcriptomic land-

scape within individual cells. One of the most exciting applications in scRNA-seq data is to find the differentially expressed (DE) gene between different cell stages or types. However, DE analysis at the single-cell level is being challenged by the cellular heterogeneity, resulting from multimodal or over-dispersed gene expression values. Most of the existing DE analysis methods are incapable of handling this heterogeneity. In this paper, we introduce the Jackknife Empirical Likelihood Ratio (JELR) test, a statistical method to address the issue. Under the null hypothesis, we show that the constructed test statistic converges to a Chi-squared distribution with degree of freedom 1. It is nonparametric and distribution-free, which is robust to model any multimodal data. Moreover, it also creates a mean-variance relationship to control the over-dispersion characteristic. Based on both simulated and real mouse scRNA-seq data, we demonstrate that JELR is an effective tool to improve the accuracy of DE analysis in heterogeneous groups and promote the study of transcriptomics in complex diseases. **Keywords:** Single-cell RNA sequencing data, Differential expression analysis, Heterogeneity, Jackknife empirical likelihood, U statistic

#### **Expectile Neural Networks for Genetic Data Analysis of Complex Diseases**

♦ *Jinghang Lin<sup>1</sup>, Xiaoran Tong<sup>1</sup>, Chenxi Li<sup>1</sup> and Qing Lu<sup>2</sup>*

<sup>1</sup>Michigan State University

<sup>2</sup>University of Florida

sttljh@gmail.com

The genetic etiologies of common diseases are highly complex and heterogeneous. Classic statistical methods, such as linear regression, have successfully identified numerous genetic variants associated with complex diseases. Nonetheless, for most complex diseases, the identified variants only account for a small proportion of heritability. Challenges remain to discover additional variants contributing to complex diseases. Expectile regression is a generalization of linear regression and provides completed information on the conditional distribution of a phenotype of interest. While expectile regression has many nice properties and holds great promise for genetic data analyses (e.g., investigating genetic variants predisposing to a high-risk population), it has been rarely used in genetic research. In this paper, we develop an expectile neural network (ENN) method for genetic data analyses of complex diseases. Similar to expectile regression, ENN provides a comprehensive view of relationships between genetic variants and disease phenotypes and can be used to discover genetic variants predisposing to sub-populations (e.g., high-risk groups). We further integrate the idea of neural networks into ENN, making it capable of capturing non-linear and non-additive genetic effects (e.g., gene-gene interactions). Through simulations, we showed that the proposed method outperformed an existing expectile regression when there exist complex relationships between genetic variants and disease phenotypes. We also applied the proposed method to the genetic data from the Study of Addiction: Genetics and Environment (SAGE), investigating the relationships of candidate genes with smoking quantity.

#### **Session 54: Statistical innovations in medicine and public health**

##### **On the Time-varying Predictive Performance of Longitudinal Biomarkers: Measure and Estimation**

*Jing Zhang<sup>1</sup>, ♦Jing Ning<sup>2</sup>, Xuelin Huang<sup>2</sup> and Ruosha Li<sup>1</sup>*

<sup>1</sup>The University of Texas Health Science Center at Houston

<sup>2</sup>The University of Texas MD Anderson Cancer Center

jing.zhang.1@uth.tmc.edu

In many biomedical studies, participants are monitored at periodic visits until the occurrence of the failure event. Biomarkers are often measured repeatedly during these visits, and such measurements can facilitate updated disease prediction. In this work, we propose a two-dimensional incident area under curve (AUC), to capture the variability due to both the biomarker assessment time and the prediction time to comprehensively quantify the predictive performance of a longitudinal biomarker. We propose a pseudo partial-likelihood to achieve consistent estimation of the AUC under two realistic scenarios of visit schedules. Variance estimation methods are designed to facilitate inferential procedures. We examine the finite-sample performance of our method through extensive simulations. The methods are applied to a study of chronic myeloid leukemia to evaluate the predictive performance of longitudinally-collected gene expression levels.

##### **New Families of Bivariate Copulas via Unit Weibull Distortion**

*Jungsywan Sepanski*

Central Michigan University

sepan1jh@cmich.edu

This paper introduces a new family of bivariate copulas constructed using a unit Weibull distortion. Existing copulas play the role of the base or initial copulas that are transformed or distorted into a new family of copulas with additional parameters, allowing more flexibility and better fit to data. We present a general form for the new bivariate copula function and its conditional and density distributions. The tail behaviors are investigated and indicate the unit Weibull distortion may result in new copulas with upper tail dependence when the base copula has no upper tail dependence. The concordance ordering and Kendall's tau are derived for the cases when the base copulas are Archimedean, such as the Clayton and Frank copulas. The Loss-ALEA data are analyzed to evaluate the performance of the proposed new families of copulas.

##### **A family of partially linear single index models for analyzing complex environmental mixtures with continuous, categorical, survival, and longitudinal health outcomes**

♦ *Yuyan Wang, Yinxiang Wu, Myeonggyun Lee, Peng Jin, Leonardo Trasande and Mengling Liu*

NYU Langone Medical Center

yuyan.wang@nyulangone.org

Studying joint effects of environmental mixtures to understand the impact of simultaneous, inter-dependent, synergistic and agnostic exposures on health outcomes remains a challenging problem. Statistical methods proposed to address this question focus on various questions, and there has been no systematic development or integrated framework that can handle many common types of outcomes. We presented a unified family of partially linear single index (PLSI) models with coherent algorithm, each consisting of a parametric linear combination of multiple exposures into a single index for practical interpretability, and a nonparametric link function for flexibility in modeling nonlinear effects and interactions. We considered PLSI linear regression and PLSI quantile regression for cross-sectional continuous outcome, PLSI generalized linear regression for cross-sectional categorical outcome, PLSI Cox proportional hazards model for survival outcome, and PLSI mixed effects model for longitudinal outcome. These models were applied to real or simulated NHANES 2003-2004 data of 800 subjects to estimate joint effects of 10 exposures on triglyceride, and they showed expected good performances.

##### **Novel empirical likelihood inference for the mean difference**

**with right-censored data**♦ *Kangni Alemjdrodo and Yichuan Zhao*Georgia State University  
kalemdjrodol@student.gsu.edu

This study focuses on comparing two means and finding a confidence interval for the difference of two means with right-censored data using the empirical likelihood (EL) method combined with the i.i.d. representation technique. In the literature, Wang and Wang (2001) proposed EL-based confidence intervals for the mean difference based on right-censored data using the synthetic data approach. However, their empirical log-likelihood ratio statistic has a scaled chi-squared distribution. To avoid the estimation of the scale parameter in constructing confidence intervals, we propose an EL method based on i.i.d. representation of Kaplan-Meier weights involved in the EL ratio. We obtain the standard chi-squared distribution. We also apply the adjusted empirical likelihood (AEL) to improve coverage accuracy for small samples. In addition, we investigate a new EL method, the mean empirical likelihood (MEL), within the framework of our study. The performances of all the EL methods are compared via extensive simulations. The proposed EL-based confidence interval has better coverage accuracy than these from Wang and Wang (2001). Finally, our findings are illustrated with a real data set.

**Semiparametric Marginal Regression Analysis for Clustered Competing Risks Data with Missing Cause of Failure**♦ *Wenxian Zhou<sup>1</sup>, Giorgos Bakoyannis<sup>1</sup>, Ying Zhang<sup>2</sup> and Constantin Yiannoutsos<sup>1</sup>*<sup>1</sup>Indiana University<sup>2</sup>University of Nebraska Medical Center  
wz11@iu.edu

Clustered competing risks data are commonly encountered in biomedical research and are often subject to missing causes of failure, intra-cluster correlation and informative cluster size. Several methods have been proposed for competing risks data with missing cause of failure or for population-averaged/cluster-specific analysis of clustered competing risks data separately. However, to the best of our knowledge, there is no general method for population-averaged inference based on clustered competing risks data with missing causes of failure. Such method is crucial for more valid inference in multicenter biomedical studies. Based on a marginal proportional cause-specific hazards model, we propose a semiparametric pseudo partial likelihood estimator for population-averaged analysis of clustered competing risks data with missing causes of failure under a missing at random assumption. We make no assumption regarding the within-cluster dependence and also allow for informative cluster size. We propose rigorous inference procedures for both regression coefficients and infinite-dimensional parameters, such as the marginal cumulative incidence function. The asymptotic properties of the proposed estimators are rigorously established using empirical process theory techniques. Simulation studies show that the proposed method performs well with finite samples and that methods ignoring the within-cluster dependence lead to invalid inferences. The proposed method is applied to clustered competing risks data from a multicenter HIV study in sub-Saharan Africa where a significant portion of causes of failure is missing. Keywords: Clustered data; Competing risks; Missing cause of failure; Informative cluster size

**A statistical methodology to characterize the relative humidity and temperature of an archaeological site**♦ *Sandra, Milena<sup>1</sup>, Ramirez<sup>1</sup> and Buelvas<sup>1</sup>*<sup>1</sup>Universidad Politécnica de Valencia  
sanrabue@doctor.upv.es

In archaeological sites, temperature differences between various minerals in block surfaces and differences in surface and substrate temperature are sources of thermal stress. Thermal and humidity stresses are important causes of micro-fractures between mineral grains in blocks. The study of microclimatic conditions surrounding archaeological sites is essential to prevent deterioration and identify eventual consequences of corrective measures. In the year 2010, the remains of the L'Almoína Museum (Valencia, Spain) displayed harmful effects on the state of conservation, as a consequence of a greenhouse effect caused by a skylight. In order to prevent the overheating of the Museum two corrective measures and changes were put in place. Furthermore, two monitoring campaigns were carried out in the years 2011 and 2013. The main goal of this study is to propose a statistical methodology to characterize the relative humidity and temperature recorded at the L'Almoína Museum during 2011 and 2013. In order to assess the effect of different corrective measures and changes implemented in the Museum.

**Session 55: Theory and Methodology for Big and Complex Data****Robust Multiple Inference for Large-Scale Multivariate Regression**♦ *Youngseok Song<sup>1</sup>, Wen Zhou<sup>1</sup> and Wen-Xin Zhou<sup>2</sup>*<sup>1</sup>Colorado State University<sup>2</sup>University of California, San Diego  
yssong@colostate.edu

Large-scale multivariate regression is a fundamental statistical tool that finds applications in a wide range of areas. This paper considers the problem of simultaneously testing a large number of general linear hypotheses, encompassing covariate-effect analysis, analysis of variance, and model comparisons. The new challenge that comes along with the overwhelmingly large number of tests is the ubiquitous presence of heavy-tailed and/or highly skewed measurement noise, which is the main reason for the failure of conventional least squares based methods. For large-scale multivariate regression, we develop a set of robust inference methods to explore data features, such as heavy tailedness and skewness, which are invisible to the scope of least squares. The new testing procedure is built on data-adaptive Huber regression, and a new covariance estimator of the regression estimate. Under mild conditions, we show that the proposed method controls the false discovery proportion and rate asymptotically. Extensive numerical experiments, along with an empirical study on quantitative linguistics, demonstrate the advantage of our proposal compared to many state-of-the-art methods

**Revisiting Convexity-Preserving Signal Recovery with the Linearly Involved GMC Penalty**♦ *Xiaoqian Liu and Eric Chi*North Carolina State University  
xliu62@ncsu.edu

The generalized minimax concave (GMC) penalty is a newly proposed regularizer, which can be used for signal recovery while maintaining the convexity of the model. Our work focuses on the linearly involved GMC penalty (composing a linear operator with the GMC penalty) and the corresponding linearly involved convexity-preserving model. First, we propose a new method to set the matrix parameter in the linearly involved GMC penalty. We transform this task into a split feasibility problem and present two

algorithms, CQ and ADMM, to solve this problem. In contrast to previous algorithms, both CQ and ADMM can be applied to any kind of linear operators. Then, we reconsider the linear involved GMC model as a saddle-point problem and use the PDHG algorithm to get the solution. Another important contribution in this work is that we provide useful guidance on the tuning parameter selection by proving some good properties of the solution path. Finally, we apply the linearly involved GMC model to 2-D image recovery. The numerical results show that the linearly involved GMC penalty can get uniformly better estimate accuracy in comparison with the standard TV regularizer.

#### Diagnosing Learning Algorithms with Super-optimal Recursive Estimators

♦ *Man Fung Leung and Kin Wai Chan*

Chinese University of Hong Kong  
heman@link.cuhk.edu.hk

Stochastic approximation algorithms have been extensively used in the fields of statistics and machine learning. Nevertheless, the total number of iterations until convergence is often unknown a priori and inferred with sequential estimates of the long-run variance (LRV). Existing LRV estimators are either statistically efficient with non-recursive algorithm or computationally efficient with higher asymptotic mean squared error (AMSE). In this paper, we develop a general framework with five principles specialized in the recursive estimation of LRV. Our main contributions lie in three different aspects: Statistically, we propose several recursive estimators with super-optimal AMSE as compared with their non-recursive counterparts. The improvements come from separating the tapering and subsampling behaviors of kernels and applying our principles. Since kernels are common in nonparametric methods, our principles can be potentially extended to other areas. To the best of our knowledge, we also derive the first sufficient condition for an estimator to be updated in constant time or space. Computationally, we introduce the concept of mini-batch estimation to improve computational efficiency beyond traditional online estimation in practice. This is important for high frequency data, which may arrive faster than executing an online algorithm. To implement our estimators, we present an automatic optimal parameters selection algorithm and a simple multivariate extension. Practically, we discuss applications to Markov chain Monte Carlo and stochastic gradient descent, which are widely used in both academia and industry. Our experiments show that the finite sample properties of our proposals match with the theoretical findings.

#### Semiparametric maximum likelihood estimation of panel count data with time-dependent covariates

♦ *Dayu Sun<sup>1</sup> and Jianguo Sun<sup>2</sup>*

<sup>1</sup>Emory University

<sup>2</sup>University of Missouri-Columbia  
commintern@hotmail.com

Panel count data appear in many clinical and observational studies, where one can only observe the counts of events between two consecutive observation times. There are two popular models of panel count data: the proportional mean/rate model assumes covariates are multiplicatively related to the mean/rate function of counts. A majority of previous studies considered maximum likelihood estimation (MLE) based on the proportional mean model because the likelihood function under the rate model involves intractable integration. However, the rate model is more realistic when covariates fluctuate over time since the non-decreasing monotonicity of the mean function may not hold. Hence, we propose a semi-parametric

MLE method under the rate model for panel count data with time-dependent covariates. The main innovation is developing an efficient Expectation-Maximization-type algorithm to overcome the computational difficulty in maximizing the likelihood under the rate model. The resulting estimators are shown to be consistent and asymptotically efficient. Monte Carlo simulation studies demonstrate that the proposed method enjoys desirable finite-sample properties. An application to a skin cancer study illustrates the proposed method in practice.

#### Regression Analysis of Multivariate Panel Count Data with Time-dependent Coefficient and Covariate Effects

♦ *Yuanyuan Guo<sup>1</sup>, Dayu Sun<sup>2</sup> and Jianguo (Tony) Sun<sup>1</sup>*

<sup>1</sup>University of Missouri - Columbia

<sup>2</sup>Emory University

yg882@mail.missouri.edu

Panel count data are prevalent in epidemiological studies, medical follow-up studies, social science, and tumorigenicity experiments. Much attention has been paid to the proportional mean model with time-constant coefficients for univariate panel count data. However, there could exist more than one type of event of interest, such as two types of tumor recurrence, leading to the multivariate panel count data. In this work, we consider multivariate panel count data with simultaneously time-dependent coefficient and covariate effects, which has not been fully investigated. Based on the conditional estimating equations method developed for time-dependent covariates, we approximate the coefficients by B-splines, hence allow both coefficients and covariates to be time-dependent. Simulation studies show that the proposed estimation procedures work well for practical situations. The methodology is applied to the China Health and Nutrition Survey (CHNS) study.

#### Smoothed empirical likelihood for the difference of two quantiles with the paired sample

♦ *Pangpang Liu and Yichuan Zhao*

Georgia State University

liu.pangpang@hotmail.com

We propose the smoothed empirical likelihood for the difference of quantiles with paired samples. The empirical likelihood for the difference of two quantiles with independent samples has been studied by some researchers. However, for many variables, we cannot ignore the correlation between the data. The correlation increases the difficulty to estimate the quantile difference. We construct two estimating equations for the difference of two quantiles and introduce a nuisance parameter in our proposed smoothed empirical likelihood. The limiting distribution of the smoothed empirical likelihood is  $\chi^2$  distribution. Simulation studies demonstrate our method is valid for the difference of quantiles. We also apply the proposed method to a real data set to illustrate the interval estimate of the quantile difference of GDP between different years.

#### Latent Network Structure Learning from High Dimensional Multivariate Point Processes

♦ *Biao Cai<sup>1</sup>, Jingfei Zhang<sup>2</sup> and Yongtao Guan<sup>1</sup>*

<sup>1</sup>Miami Herbert Business School

<sup>2</sup>Miami Herbert School

bcai@bus.miami.edu

Learning the latent network structure from large scale multivariate point process data is an important task in a wide range of scientific and business applications. For instance, we might wish to estimate the neuronal functional connectivity network based on spiking times recorded from a collection of neurons. To characterize the complex processes underlying the observed data, we propose a new and flex-

ible class of nonstationary Hawkes processes that allow both excitatory and inhibitory effects. We estimate the latent network structure using an efficient sparse least squares estimation approach. Using a thinning representation, we establish concentration inequalities for the first and second order statistics of the proposed Hawkes process. Such theoretical results enable us to establish the non-asymptotic error bound and the selection consistency of the estimated parameters. Furthermore, we describe a penalized least squares based statistic for testing if the background intensity is constant in time. We demonstrate the efficacy of our proposed method through simulation studies and an application to a neuron spike train data set.

## Session 56: Keynote speech

### Use of Real World Healthcare Data to Accelerate Vaccine Development in the Post COVID Era

*Josh Chen*

Sanofi Pasteur

[Josh.Chen@sanofi.com](mailto:Josh.Chen@sanofi.com)

Human vaccine research and development is a lengthy, risky and expensive process which typically takes 10-15 years from discovery to approval. Lessons learned from the current collaborative efforts to develop safe and effective COVID-19 vaccines within 12-18 months support the aspiration that it is possible to accelerate vaccine development using innovative approaches. Before the COVID pandemic, there had been strong interest in the potential use of real-world evidence for regulatory purposes. The COVID pandemic will further catalyzes digital transformation and advancement of information technology infrastructure and as a result, vast increase in high quality real world data pertaining to patient health and healthcare delivery. In this talk, we will advocate use of real world data from healthcare systems, including electronic health records (EHRs), medical claims and billing data, and patient registries, to generate fit-for-purpose real world evidence in support of the safety and effectiveness of an experimental vaccine for regulatory decisions.

## Session 57: Statistical Applications of Extreme Value Theory

### Semi-parametric estimation for multivariate extremes

♦*John Nolan<sup>1</sup>, Anne-Laure Fougères<sup>2</sup> and Cecile Mercadier<sup>2</sup>*

<sup>1</sup>American University

<sup>2</sup>University of Lyon

[jpnolan@american.edu](mailto:jpnolan@american.edu)

We present a new way to estimate multivariate extreme value distributions (MVEVD) from data using max projections. The approach works in any dimension, though computation time increases quickly as dimension increases. The procedure requires tools from computational geometry and multivariate integration techniques. An R package `mvevd` is being developed to implement the method for several semi-parametric classes of MVEVDs: discrete angular measure, generalized logistic, piecewise linear angular measures, and Dirichlet mixture models.

### All block maxima method for estimating the extreme value index

*Jochem Oorschot and ♦Chen Zhou*

Erasmus University Rotterdam

[zhou@ese.eur.nl](mailto:zhou@ese.eur.nl)

The block maxima (BM) approach in extreme value analysis is a sample of block maxima to the Generalized Extreme Value (GEV)

distribution. We consider all potential blocks from a sample, which leads to the All Block Maxima (ABM) estimator. Different from existing estimators based on the BM approach, the ABM estimator is permutation invariant. We show the asymptotic behavior of the ABM estimator, which has the lowest asymptotic variance among all estimators using the BM approach. Simulation studies justify our asymptotic theories. A key step in establishing the asymptotic theory for the ABM estimator is to obtain asymptotic expansions for the tail empirical process based on higher order statistics with weights.

### Dynamic Bivariate Peak over Threshold Model for Joint Tail Risk Dynamics of Financial Markets

*Zifeng Zhao*

University of Notre Dame

[zzhao2@nd.edu](mailto:zzhao2@nd.edu)

We propose a novel dynamic bivariate peak over threshold (PoT) model to study the time-varying behavior of joint tail risk in financial markets. The proposed framework provides simultaneous modeling for dynamics of marginal and joint tail risk, and generalizes the existing tail risk literature from univariate dimension to multivariate dimension. We introduce a natural and interpretable tail connectedness measure and examine the dynamics of joint tail behavior of global stock markets: empirical evidence suggests markets from the same continent have time-varying and high-level joint tail risk, and tail connectedness increases during periods of crisis. We further enrich the tail risk literature by developing a novel portfolio optimization procedure based on bivariate joint tail risk minimization, which gives promising risk-rewarding performance in backtesting.

### A Preferential Attachment Model with Poisson Growth

♦*Tiandong Wang<sup>1</sup> and Sidney Resnick<sup>2</sup>*

<sup>1</sup>Texas A&M University

<sup>2</sup>Cornell University

[twang@stat.tamu.edu](mailto:twang@stat.tamu.edu)

Recent studies on the evolution of regional social networks reveal that the number of edges created per day follows a non-homogeneous Poisson process with constant rates within a day but varying rates from day to day. Also, for real networks, it is often possible to have coarse timestamp information. We then introduce a variant of the traditional preferential attachment model by taking into account both the Poisson growth and the low resolution in the timestamp information for a temporal network. Analytical properties of this new model are then studied and we fit the new model to a couple of different real datasets.

## Session 58: Variable selection with complex lifetime data

### Simultaneously Variable Selection and Estimation for Interval-Censored Failure Time Data

*Jianguo Sun*

University of Missouri

[sunj@missouri.edu](mailto:sunj@missouri.edu)

Variable selection is a common task in many fields and also a hot topic for the analysis of high-dimensional data. Correspondingly, many methods have been developed and among them, a general type of procedures is the penalized approach. In this talk, we will discuss variable selection when one faces interval-censored failure time data, a general type of failure time data that can occur in many areas including demographical studies, economic studies, medical studies and social sciences. For the problem, a new penalized procedure will be presented and discussed.

### Learning survival from EMR/EHR data to estimate treatment effects using high dimensional claims codes

Ronghui Xu

UC San Diego  
rxu@ucsd.edu

Our work was motivated by the analysis projects using the linked US SEER-Medicare database to study treatment effects in men of age 65 years or older who were diagnosed with prostate cancer. Such data sets contain up to 100,000 human subjects and over 20,000 claim codes. The data were obviously not randomized with regard to the treatment of interest, for example, radical prostatectomy versus conservative treatment. Informed by previous instrumental variable (IV) analysis, we know that confounding most likely exists beyond the commonly captured clinical variables in the database, and meanwhile the high dimensional claims codes have been shown to contain rich information about the patients' survival. Hence we aim to incorporate the high dimensional claims codes into the estimation of the treatment effect. The orthogonal score method is one that can be used for treatment effect estimation and inference despite the bias induced by regularization under the high dimensional hazards outcome model and the high dimensional treatment model. In addition, we show that with cross-fitting the approach has rate doubly-robust property in high dimensions.

### Model Large-Scale Survival Data with Time-Varying Effects via a Minorization-Maximization Steepest Ascent Algorithm

◆Zhi (Kevin) He, Ji Zhu, Jian Kang and Yi Li

University of Michigan  
kevinhe@umich.edu

Time-varying effect survival models present a flexible tool for modeling risk factors with dynamic effects. Fitting the model, however, carries much computational burden, especially when the sample size or the number of predictors is large. For example, fitting a time-varying effect model on a national kidney transplant data set with about 300,000 subjects and 100 predictors may defy any existing statistical methods or software. To relieve the cumbersome computation, we propose a Minorization-Maximization steepest ascent procedure, which leverages the block structure formed by the basis expansions for each coefficient function, and iteratively updates the optimal block-wise search direction, along which the increase of the partial likelihood is maximized. We show that the updated estimates always increase the partial likelihood until convergence. We further propose a score test to examine whether the effects are indeed time-varying. We evaluate the utility of the proposed method via simulations, and apply the method to analyze the national kidney transplant data and illustrate the time-varying effects of various risk factors.

### Variable selection for joint models with time-varying coefficients

Yujing Xie<sup>1</sup>, ◆Zangdong He<sup>2</sup>, Wanzhu Tu<sup>3</sup> and Zhangsheng Yu<sup>1</sup>

<sup>1</sup>Shanghai Jiao Tong University

<sup>2</sup>GlaxoSmithKline

<sup>3</sup>Indiana University  
zangdong.x.he@gsk.com

Many clinical studies collect longitudinal and survival data concurrently. Joint models combining these two types of outcomes through shared random effects are frequently used in practical data analysis. The standard joint models assume that the coefficients for the longitudinal and survival components are time-invariant. In many applications, the assumption is overly restrictive. In this research, we extend the standard joint model to include time-varying coefficients, in both longitudinal and survival components, and we

present a data-driven method for variable selection. Specifically, we use a B-spline decomposition and penalized likelihood with adaptive group LASSO to select the relevant independent variables and to distinguish the time-varying and time-invariant effects for the two model components. We use Gaussian-Legendre and Gaussian-Hermite quadratures to approximate the integrals in the absence of closed-form solutions. Simulation studies show good selection and estimation performance. Finally, we use the proposed procedure to analyze data generated by a study of primary biliary cirrhosis.

### Session 59: Recent advances in statistical methods for big biomedical data integration

#### Outcome-guided Sparse K-means for Disease Subtype Discovery via Integrating Phenotypic Data with High-dimensional Transcriptomic Data

Lingsong Meng<sup>1</sup>, Dorina Avram<sup>2</sup>, George Tseng<sup>3</sup> and ◆Zhiguang Huo<sup>1</sup>

<sup>1</sup>Department of Biostatistics, University of Florida

<sup>2</sup>Department of Immunology, H. Lee Moffitt Cancer Center and Research Institute

<sup>3</sup>Department of Biostatistics, University of Pittsburgh  
zhuo@ufl.edu

The discovery of disease subtypes is an essential step for developing precision medicine, and disease subtyping via omics data has become a popular approach. While promising, subtypes obtained from current approaches are not necessarily associated with clinical outcomes. With the rich clinical data along with the omics data in modern epidemiology cohorts, it is urgent to develop an outcome-guided clustering algorithm to fully integrate the phenotypic data with the high-dimensional omics data. Hence, we extended a sparse K-means method to an outcome-guided sparse K-means (GuidedSparseKmeans) method, which incorporated a phenotypic variable from the clinical dataset to guide gene selections from the high-dimensional omics data. We demonstrated the superior performance of the GuidedSparseKmeans by comparing with existing clustering methods in simulations and applications of high-dimensional transcriptomic data of breast cancer and Alzheimer's disease.

#### Multiple testing correction for multivariate-multivariate association analysis: with application to an imaging-genetics study

◆Shuo Chen<sup>1</sup> and Qiong Wu<sup>2</sup>

<sup>1</sup>University of Maryland

<sup>2</sup>University of Maryland, College Park  
shuochen@som.umaryland.edu

We consider a complex multiple comparison problem regarding multivariate-multivariate association analysis. Our method development is motivated by the recent advent of imaging-genetics technology, facilitating the simultaneous collection of massive genetics and brain imaging data. The number of simultaneous tests between two high-dimensional data can easily rise to hundreds of billions. Due to the limited sample size, most studies are underpowered when applying traditional multiple testing correction methods (e.g., FDR). To address this challenge, we propose a new graph and combinatorics based multiple correction method to identify the set-wise association (e.g., a set of alleles are associated with a set of imaging features). We control the family-wise error rate at the set pair level. Our approach can drastically improve statistical power by capitalizing on the intrinsically organized pattern of the multivariate-multivariate association. We perform extensive simulation studies

and compare our method with existing methods. The results show that our method provides accurate multi-to-multi statistical inference with increased power and reduced false-positive error rate. We also apply this new approach to an imaging-genetics data set and identify biologically meaningful brain-area-genetic-loci association sets.

#### **Adaptive integration of testing results from multiple-trait genome-wide association studies**

♦ *Chi Song and Qiaolan Deng*

Ohio State University  
song.1188@osu.edu

In genome-wide association studies (GWASs), there is an increasing need for detecting the associations between a genetic variant and multiple traits. In studies of complex diseases, it is common to measure several potentially correlated traits in a single GWAS. Despite the multivariate nature of the studies, single-trait-based methods remain the most widely-adopted analysis procedure, owing to their simplicity for studies with multiple traits as their outcome. However, the association between a genetic variant and a single trait sometimes can be weak, and ignoring the actual correlation among traits may lose power. On the contrary, multiple-trait analysis, a method analyzes a group of traits simultaneously, has been proven to be more powerful by incorporating information from the correlated traits. Although existing methods have been developed for multiple traits, several drawbacks limit their wide application in GWASs. First, many existing methods can only process continuous traits and fail to allow for binary traits which are ubiquitous in the real-world problems. Second, as shown in our simulation study, the performance of many existing methods is unstable under different scenarios where the correlation among traits and the signal proportion vary. In this talk, we propose a multiple-trait adaptive Fisher's (MTAF) method to test associations between a genetic variant and multiple traits at once, by adaptively integrating evidence from each trait. The proposed method can accommodate both continuous and binary traits and it has reliable performance under various scenarios. Using a simulation study, we compared our proposed method with several existing methods and demonstrated its competitiveness in terms of type I error control and statistical power. By applying the method to the Study of Addiction: Genetics and Environment (SAGE) dataset, we successfully identified several genes associated with substance dependence.

#### **Combining p-values under arbitrary dependency structure in heavy-tailed distributions**

*Yusi Fang<sup>1</sup>, Chung Chang<sup>2</sup>, Yongseok Park<sup>1</sup> and ♦George Tseng<sup>1</sup>*

<sup>1</sup>University of Pittsburgh  
<sup>2</sup>National Sun Yat-Sen University  
ctseng@pitt.edu

The issue of combining individual p-values to aggregate multiple small effects is prevalent in many scientific investigations and is a long-standing statistical topic. Many classical methods are designed for combining independent and frequent signals in a traditional meta-analysis sense using sum of transformed p-values with transformation of light-tailed distributions, in which Fisher and Stouffer methods are most well-known. Since early 2000, advances in big data promoted methods to aggregate independent, sparse and weak signals, such as renowned higher criticism and Berk-Jones tests. Recently, Liu and Xie (2020) and Wilson (2019) independently proposed Cauchy and harmonic mean combination tests to robustly combine p-values under arbitrary dependency structure, where a notable application is to combine p-values of multiple correlated SNPs

in a SNP-set in genome-wide association studies. The proposed tests are transformation of heavy-tailed distributions for improved power with sparse signal. It calls for a natural question to investigate heavy-tailed distribution transformation, to understand connection among existing methods and explore the necessary and sufficient condition for a method to possess robustness to dependence structure. In this paper, we investigate the regularly-varying distribution, a rich family of heavy-tailed distribution including Pareto distribution as a special case. We show that only an equivalent class of Cauchy and harmonic mean tests have the robustness. We also show an issue caused by large negative penalty in the Cauchy method and propose a simple modification. Finally, we present simulations and apply to a neuroticism GWAS application to verify the discovered theoretical insights.

#### **Session 60: Complex data analysis in business, economics, and industry**

##### **On Model Selection for ARFIMA and GARCH Processes**

*Ngai Hang Chan<sup>1</sup>, ♦Kun Chen<sup>2</sup>, Hsueh-Han Huang<sup>3</sup> and Ching-Kang Ing<sup>3</sup>*

<sup>1</sup>Chinese University of Hong Kong

<sup>2</sup>Southwestern University of Finance and Economics

<sup>3</sup>National Tsing Hua University  
arsenalplay@163.com

It was shown in Beran et al. (1998) that Bayesian information criterion (BIC) and Hannan-Quinn information criterion (HQ) are order selection consistent in fractional autoregressive processes with constraints on the memory parameter  $d$ . The estimation and prediction problems in autoregressive fractionally integrated moving average (ARFIMA) models with  $d \in \mathbb{R}$  have also been considered by Hualde and Robinson (2011) and Chan et al. (2013), respectively. However, order selection consistency in the ARFIMA model still remains unsolved. In this paper, we fill this gap by proposing a novel model selection procedure and proving the desired selection consistency result without constraints on  $d$ . The result provides a unified treatment of fractional and non-fractional, stationary and integrated non-stationary ARMA models. Also, we applied the result to select the order for generalized autoregressive conditional heteroskedasticity (GARCH) models. Numerical analysis is conducted to illustrate our theoretical findings.

##### **Testing for change points in heavy-tailed time series—A trimmed CUSUM approach**

*She Rui<sup>1</sup> and ♦Ling Shi<sup>2</sup>*

<sup>1</sup>Southwestern University of Finance and Economics

<sup>2</sup>HKUST  
maling@ust.hk

It is well-known how to test the change-point in heavy-tailed time series is greatly open. In this article, we propose a trimmed CUSUM (cumulative sum) approach to solve the problem. We first study the trimmed CUSUM process and give its limiting properties under the null hypothesis and the alternative. Based on this fundamental result, the theories for Kolmogorov-Smirnov test and Self-normalized(SN) tests are established. Since our assumption is very weak, the proposed tests are not only flexible to linear time series but also nonlinear time series, such as TAR and G-GARCH. In addition, we further consider the multiple change-point alternative, where we derive some novel results to the powers which further reveal the essential difference between the single change-point SN test and the multiple. Simulation studies are carried out to assess



the performance of our procedure to test the stability in mean and volatility. Two real examples in financial markets are illustrated. All these empirical evidences show that the proposed procedure has an encouraging performance in detecting the structural stability of a heavy-tailed time series.

#### A unified approach to bias approximations

♦ *Ruby Chiu-Hsing Weng*<sup>1</sup> and *Derek Stephen Coad*<sup>2</sup>

<sup>1</sup>National Chengchi University

<sup>2</sup>Queen Mary, University of London  
chwend@nccu.edu.tw

Bias approximation has played an important role in statistical inference, and numerous bias calculation techniques have been proposed under different contexts. We provide a unified approach to approximating the bias of the maximum likelihood estimator and the l2 penalized likelihood estimator for both linear and nonlinear models, where the design variables are allowed to be random and the sample size can be a stopping time. The proposed method is justified by very weak approximations. The accuracy of the derived bias formulas is assessed by simulation for several examples. The bias of the l1 penalized estimator will be briefly discussed.

#### Model Averaging for High-dimensional Linear Regression Models with Dependent Observations

♦ *Ting-Hung Yu*<sup>1</sup>, *Ching-Kang Ing*<sup>2</sup> and *Henghsiu Tsai*<sup>3</sup>

<sup>1</sup>University of Iowa, U.S.A.

<sup>2</sup>National Tsing Hua University

<sup>3</sup>Academia Sinica  
ting-hung-yu@uiowa.edu

We introduce the orthogonal greedy algorithm (OGA) to screen out the nested set of signal variables under a high-dimensional linear regression framework with dependent observations. To gain the prediction performance, we propose the high-dimensional Mallows model averaging (HDMMA) criteria to determine the weight for averaging these nested high-dimensional linear regression models. We further analyze rates of convergence of prediction error for the averaging model under different sparsity conditions. Our contribution has three folds. First, we show that our procedure, named OGA+HDMMA, can achieve optimal convergence rates of prediction error discussed in Ing (2019). Second, we use simulation to show that the out-sample prediction of OGA+HDMMA can outperform the MCV method proposed in Ando and Li(2014) when the covariates are highly correlated or contain time-series effects. Third, the out-sample prediction of OGA+HDMMA performs comparably or even better than many well-known high-dimensional variables selection methods in some scenarios. Keywords: High-dimensional Mallows model averaging, orthogonal greedy algorithm, sparsity conditions, time series, high-dimensional linear regression models, optimal rates of convergence.

### Session 61: Bayesian Analysis of Complex and High Dimensional Data

#### Spatio-Temporal Additive Regression Model Selection for Urban Water Demand

*Hunter Merrill*<sup>1</sup>, *Xueying Tang*<sup>2</sup> and ♦ *Nikolay Bliznyuk*<sup>1</sup>

<sup>1</sup>University of Florida

<sup>2</sup>University of Arizona  
nbliznyuk@ufl.edu

Understanding the factors influencing urban water use is critical for meeting demand and conserving resources. To analyze the relationships between urban household-level water demand and poten-

tial drivers, we develop a method for Bayesian variable selection in partially linear additive regression models, particularly suited for high-dimensional spatio-temporally dependent data. Our approach combines a spike-and-slab prior distribution with a modified version of the Bayesian group lasso to simultaneously perform selection of null, linear, and nonlinear models and to penalize regression splines to prevent overfitting. We investigate the effectiveness of the proposed method through a simulation study and provide comparisons with existing methods. We illustrate the methodology on a case study to estimate and quantify uncertainty of the associations between several environmental and demographic predictors and spatio-temporally varying household-level urban water demand in Tampa, FL.

#### Joint Bayesian Analysis of Multiple Response-Types Using the Hierarchical Generalized Transformation Model

*Jonathan Bradley*

Florida State University  
jrbradley@fsu.edu

Suppose we observed data consisting of multiple response-types (e.g., continuous, count-valued, Bernoulli trials, etc.), which are distributed from more than one class of distributions. We refer to these types of data as “multiple response-type” datasets. The goal of this article is to introduce a reasonable method that “converts” a Bayesian statistical model for continuous responses into a Bayesian model for multiple response-type datasets. To do this, we consider a transformation of the multiple response-type data. What is unique with our strategy is that we treat the transformations as unknown and use a Bayesian approach to model this uncertainty. The implementation of our Bayesian approach to unknown transformations is straightforward, and involves two steps. The first step produces posterior replicates of the transformed multiple response-type data from a latent conjugate multivariate (LCM) model. The second step involves generating values from the posterior distribution implied by the preferred model. We demonstrate the flexibility of our model through an application to Bayesian additive regression trees (BART) and a spatio-temporal mixed effects (SME) model. We provide a thorough joint multiple response-type spatio-temporal analysis of coronavirus disease 2019 (COVID-19) cases, the adjusted closing price of the Dow Jones Industrial (DJI), and Google Trends data.

#### Joint Bayesian Variable and DAG Selection Consistency for High-dimensional Regression Models with Network-structured Covariates

♦ *Xuan Cao*<sup>1</sup> and *Kyoungjae Lee*<sup>2</sup>

<sup>1</sup>University of Cincinnati

<sup>2</sup>Inha University  
caox4@ucmail.uc.edu

We consider the joint sparse estimation of regression coefficients and the covariance matrix for covariates in a high-dimensional regression model, where the predictors are both relevant to a response variable of interest and functionally related to one another via a Gaussian directed acyclic graph (DAG) model. Gaussian DAG models introduce sparsity in the Cholesky factor of the inverse covariance matrix, and the sparsity pattern in turn corresponds to specific conditional independence assumptions on the underlying predictors. A variety of methods have been developed in recent years for Bayesian inference in identifying such network-structured predictors in regression setting, yet crucial sparsity selection properties for these models have not been thoroughly investigated. In this paper, we consider a hierarchical model with spike and slab priors on the regression coefficients and a flexible and general class of

DAG-Wishart distributions with multiple shape parameters on the Cholesky factors of the inverse covariance matrix. Under mild regularity assumptions, we establish the joint selection consistency for both the variable and the underlying DAG of the covariates when the dimension of predictors is allowed to grow much larger than the sample size. We demonstrate that our method outperforms existing methods in selecting network-structured predictors in several simulation settings.

### Shrinkage on Simplex : Sparsity-Inducing Priors for Compositional Data

*Jyotishka Datta*

University of Arkansas  
jd033@uark.edu

Global-local shrinkage priors have been established as the current state-of-the art Bayesian tool for sparse signal detection leading to a huge literature proposing elaborate shrinkage priors for real-valued parameters. However, there has been limited consideration of discrete data structures including sparse compositional data, routinely occurring in microbiomics. I will discuss two methodological challenges. First, the Dirichlet is highly inflexible as a shrinkage prior for high-dimensional probabilities for its inability to adapt to an arbitrary level of sparsity. We address this gap by proposing the Sparse Generalized Dirichlet distribution, specially designed to enable scaling to data with many categories. A related problem is associating the compositional response data with environmental or clinical predictors. I will develop Bayesian variable selection strategies using global-local shrinkage priors for detecting significant associations between available covariates and taxonomic abundance tables. I will provide some theoretical support for the proposed methods and show improved performance in several simulation settings and application to microbiome data.

### Session 62: New statistical methods for machine learning on big data

#### Adversarial Machine Learning – Game Theoretic Approach and Adversarial Attack Against Deep Neural Networks

*Bowei Xi*

Purdue University  
xbw@purdue.edu

As more and more security data are collected, machine learning techniques become an essential tool for real-world security applications. One of the most important differences between cyber security and many other applications is the existence of malicious adversaries that actively adapt their behavior to make the existing learning models ineffective. Unfortunately, traditional learning techniques are insufficient to handle such adversarial problems directly. The adversaries adapt to the defender's reactions, and learning algorithms constructed based on the current training dataset degrades quickly. Based on a game theoretic framework to model the sequential actions of the adversaries and the defender, we develop adversarial classification and adversarial clustering methods to defend against active adversaries. An adversarial attack against deep neural networks is introduced in this talk too.

#### A universal event detection framework for Neuropixels data

♦*Hao Chen*<sup>1</sup>, *Shizhe Chen*<sup>1</sup> and *Xinyi Deng*<sup>2</sup>

<sup>1</sup>University of California, Davis

<sup>2</sup>Beijing University of Technology  
hxchen@ucdavis.edu

Neuropixels probes present exciting new opportunities for neuroscience, but such large-scale high-density recordings also introduce unprecedented challenges in data analysis. Neuropixels data usually consist of hundreds or thousands of long stretches of sequential spiking activities that evolve non-stationarily over time and are often governed by complex, unknown dynamics. Extracting meaningful information from the Neuropixels recordings is a non-trivial task. Here we introduce a general-purpose, graph-based statistical framework that, without imposing any parametric assumptions, detects points in time at which population spiking activity exhibits simultaneous changes as well as changes that only occur in a subset of the neural population, referred to as “change-points”. The sequence of change-point events can be interpreted as a footprint of neural population activities, which allows us to relate behavior to simultaneously recorded high-dimensional neural activities across multiple brain regions. We demonstrate the effectiveness of our method with an analysis of Neuropixels recordings during spontaneous behavior of an awake mouse in darkness. We observe that change-point dynamics in some brain regions display biologically interesting patterns that hint at functional pathways, as well as temporally-precise coordination with behavioral dynamics. We hypothesize that neural activities underlying spontaneous behavior, though distributed brainwide, show evidences for network modularity. Moreover, we envision the proposed framework to be a useful off-the-shelf analysis tool to the neuroscience community as new electrophysiological recording techniques continue to drive an explosive proliferation in the number and size of data sets.

#### Improved double robust approach for precision medicine

*Lingsong Zhang*

Purdue University

lingsong@purdue.edu

An energy-distance based approach is proposed in this paper to estimate the treatment effect in precision medicine setting. The new approach avoids the estimation of the typical propensity score, but can also achieve a new type of double robustness. Theoretical justification and examples will be used to show the usefulness of this new approach. This is a joint work with Zeyu Zhang and Haoda Fu.

#### A new clustering algorithm for assigning cells to known cell types according to marker genes

*Hongyu Guo* and ♦*Jun Li*

University of Notre Dame

jun.li@nd.edu

On single-cell RNA-sequencing data, we consider the problem of assigning cells to known cell types, assuming that the identities of cell-type-specific marker genes are given but their exact expression levels are unavailable, that is, without using a reference dataset. Based on an observation that the expected over-expression of marker genes is often absent in a nonnegligible proportion of cells, we develop a new clustering method called scSorter. scSorter allows marker genes to express at a low level and borrows information from the expression of non-marker genes. On both simulated and real data, scSorter shows much higher power compared to existing methods.

### Session 63: Innovative statistical methods for optimal treatment selection and clinical trial design with historical data

#### Machine Learning of Non-Randomized Control Studies for Causal Inference

♦Xiaolong Luo, Mingyu Li, Jing Gong, Marie-Laure Casadebaig, Daniel Li and Mike Branson

Bristol Meyers Squibb  
xiaolongluo1979@gmail.com

Non-randomized control studies including Real World Evidence, Synthetic Control and data mining are getting increased attention and become important in biomarker identification, regulatory approval and post marketing commitment for pharmaceutical product development. However, both data quality and existing methodology have been factors leading to concern for their applicability. In this talk, we propose a new way to overcome some methodological challenge by combining modern data science technology of deep learning and traditional counting process technique. It uses a counterfactual framework for survival data and applies machine learning to provide robust estimand that incorporates all intercurrent events. We apply the method to analysis of the overall survival endpoint from the integrated data from two oncology clinical trials that include substantial non-randomized subsequent anticancer therapies. Estimand of overall survival for each treatment policy of initial treatment and subsequent therapy will be calculated. Comparison between treatment policies will be used as causal inference to assess optimal efficacy with respect to sequential combination of study drug and subsequent therapy.

#### Evaluation of different analytic strategies for estimating optimal treatment regimens for time-to-event outcomes in observational data

♦Ilya Lipkovich, Zbigniew Kadziola, Bohdana Ratitch, Zhanglin Cui and Douglas Faries

Eli Lilly and Company  
ilya.lipkovich@lilly.com

In this presentation, we provide an overview of existing machine learning methods for evaluating individualized treatment regimens (ITR) optimizing time-to-event outcomes (restricted mean survival time, RMST) when using observational data with non-randomized treatment assignment (such as Electronic Health Records). We present a simulation study closely mimicking an observational dataset under various scenarios including different degree of alignment between observed regimens (reflecting actual prescribing practices) and optimal ITR. The simulation results include performance characteristics of the candidate methods (including Random Forests, Gradient Boosting and Outcome Weighted Learning) in terms of their ability to recover the true optimal treatment regimens, as well as operating characteristics (Bias, MSE) of various empirical measures of gain in RMST resulting from applying estimated optimal ITR comparing to following actual prescribing practices.

#### Using Real-World Evidence at FDA/CBER

Jiang (Jessica) Hu

NA  
jiang.hu@fda.hhs.gov

Based on the 21st Century Cures Act (Cure Act), FDA has created a framework to evaluate the feasibility of using real world evidence in supporting the approval of new indication or post-approval study requirement. This presentation introduces the classification of real world data and real world evidence, FDA's real world evidence pro-

gram, CDISC RWD Connect as an CDISC extension to real world data and real world evidence. It also discusses how to combine the innovative statistical methodologies, CDISC data standard extension, and FDA/CBER daily regulatory review and research work.

### Session 64: Statistical method advancement for analyzing omics data

#### Robust partial reference-free cell composition estimation from tissue expression

♦Ziyi Li<sup>1</sup>, Zhenxing Guo<sup>1</sup>, Ying Cheng<sup>2</sup>, Peng Jin<sup>1</sup> and Hao Wu<sup>1</sup>

<sup>1</sup>Emory University

<sup>2</sup>Yunnan University  
ziyi.li@emory.edu

High cost, intensive labor requirements and technical limitations hinder the cell composition quantification using cell sorting or single-cell technology. As alternatives, reference-based deconvolution algorithms are limited if no appropriate reference panel from purified tissues is available, while reference-free deconvolution algorithms are suffered from low accuracy and difficulty in cell type label assignment. Here, we introduce TOAST-P and TOAST/+P, two partial reference-free algorithms for estimating cell composition of heterogeneous tissues from their gene expression profiles. Guided by prior information obtainable from various sources, the proposed method can capture the cell composition significantly better than existing methods in extensive simulation studies and real data analyses. We evaluate the markers obtained from different high-throughput modalities and existing repositories, confirming the proposed method performs better than existing methods wherever the prior knowledge types is obtained. Finally, the analyses of two Alzheimer's disease datasets show consistency of the proposed method with cell compositions from single cell technology and existing knowledge about the disease.

#### Joint model of temporal microbiome and risk of outcomes at matched time

♦Qian Li<sup>1</sup>, Kendra Vehik<sup>1</sup>, Jeffery Krischer<sup>1</sup> and Yijuan Hu<sup>2</sup>

<sup>1</sup>University of South Florida

<sup>2</sup>Emory University  
qian.li@epi.usf.edu

Temporal gut microbiome has been found associated with different types of exposures, growth phase, and disease onset. Microbiome composition at multiple time points are frequently used to identify microbes predictive of endpoint event status in longitudinal studies. However, sample availability and metagenomic sequencing cost may restrict microbiome profiling and analysis to subjects whose outcomes (i.e. cases and controls) were defined and matched at cases' endpoint age. We proposed a joint modeling approach to analyze whether the mean (intercept) or rate-of-change over time (slope) of microbe relative abundance contributes to the risk of time-matched outcomes. The relative abundance of a microbe was specified by Zero-Inflated Beta generalized linear mixed effect sub-model, with both prevalence and non-zero abundance predicted by multiple factors or covariates. The risk of endpoint outcome was modeled by logistic regression. We used subject-level random effect to account for the correlation between temporal microbiome measurements, which was also included in the sub-model of logistic regression. Age of matched outcomes was incorporated in the logistic regression sub-model as a weight, adjusting for matched time-to-event effect. Simulation studies and real data analysis showed

that the joint model of temporal microbiome composition and time-matched outcomes outperformed a two-stage approach.

### **Integrative Analysis of Multi-Omic Data via Sparse Multiple Co-Inertia Analysis**

*Eun Jeong Min and Qi Long*

University of Pennsylvania  
qlong@penmedicine.upenn.edu

Multiple co-inertia analysis (mCIA) is a multivariate analysis method that can assess relationships and trends in multiple datasets. Recently it has been used for an integrative analysis of multiple high-dimensional -omics datasets. However, the estimated loading vectors from the existing mCIA method are non-sparse, which presents challenges for interpreting analysis results. We propose two new mCIA methods: 1) a sparse mCIA (smCIA) method that produces sparse loading estimates and 2) a structured sparse mCIA (ssmCIA) method that further enables the incorporation of structural information among variables such as those from functional genomics. The two proposed methods achieve simultaneous model estimation and feature selection, and yield analysis results that are more interpretable than the existing mCIA. Our simulation studies demonstrate the superior performance of the smCIA and the ssmCIA methods compared to the existing mCIA. We also apply our methods to integrative analysis of -omics data from a cancer study.

### **WEVar: a novel statistical learning framework for predicting noncoding regulatory variants**

*Li Chen*

Indiana University School of Medicine  
chen61@iu.edu

Understanding the functional consequence of noncoding variants is of great interest. Though genome-wide association studies (GWAS) or quantitative trait locus (QTL) analyses have identified variants associated with traits or molecular phenotypes, most of them are located in the noncoding regions, making the identification of casual variants a particular challenge. Existing computational approaches developed for prioritizing noncoding variants produce inconsistent and even conflicting results. To address these challenges, we propose a novel statistical learning framework, which directly integrates the precomputed functional scores from representative scoring methods. It will maximize the usage of integrated methods by automatically learning the relative contribution of each method and produce an ensemble score as the final prediction. The framework consists of two modes. The first “context-free” mode is trained using curated causal regulatory variants from a wide range of context and is applicable to predict noncoding variants of unknown and diverse context. The second “context-dependent” mode further improves the prediction when the training and testing variants are from the same context. By evaluating the framework via both simulation and empirical studies, we demonstrate that it outperforms integrated scoring methods and the ensemble score successfully prioritizes experimentally validated regulatory variants in multiple risk loci.

## **Session 65: Statistical learning with complex data structure**

### **Quantile Trend Filtering**

*Eric Chi<sup>1</sup>, Halley Brantley<sup>2</sup> and Joseph Guinness<sup>3</sup>*

<sup>1</sup>North Carolina State University

<sup>2</sup>United Health Group

<sup>3</sup>Cornell University  
eric\_chi@ncsu.edu

We address the problem of estimating smoothly varying baseline trends in time series data. This problem arises in a wide range of fields, including chemistry, macroeconomics and medicine; however, our study is motivated by the analysis of data from low cost air quality sensors. Our methods extend the quantile trend filtering framework to enable the estimation of multiple quantile trends simultaneously while ensuring that the quantiles do not cross. To handle the computational challenge posed by very long time series, we propose a parallelizable alternating direction method of multipliers (ADMM) algorithm. The ADMM algorithm enables the estimation of trends in a piecewise manner, both reducing the computation time and extending the limits of the method to larger data sizes. We also address smoothing parameter selection and propose a modified criterion based on the extended Bayesian information criterion. Through simulation studies and our motivating application to low cost air quality sensor data, we demonstrate that our model provides better quantile trend estimates than existing methods and improves signal classification of low-cost air quality sensor output.

### **Network Inference from Grouped Observations**

*Yunpeng Zhao<sup>1</sup>, Peter Bickel<sup>2</sup> and Charles Weko<sup>3</sup>*

<sup>1</sup>Arizona State University

<sup>2</sup>University of California, Berkeley

<sup>3</sup>US Army

yunpeng.zhao@asu.edu

Statistical network analysis typically deals with inference concerning various parameters of an observed network. In several applications, especially those from social sciences, behavioral information concerning groups of subjects are observed. Over the past century a number of descriptive statistics have been developed to infer network structure from such data. However, these measures lack a generating mechanism that links the inferred network structure to the observed groups. In this talk, we present a model-based approach called the hub model, which belongs to a family of Bernoulli mixture models. We further present theoretical results on model identifiability, a notoriously difficult problem in Bernoulli mixture models, and estimation consistency.

### **A Flexible Latent Space Model for Multilayer Networks**

*Xuefei Zhang, Songkai Xue and Ji Zhu*

University of Michigan

xfzhang@umich.edu

Entities often interact with each other through multiple types of relations, which can be represented as multilayer networks. Multilayer networks among the same set of nodes usually share common structures, while each layer can also possess its distinct node connecting behaviors. This paper proposes a flexible latent space model for multilayer networks for the purpose of capturing such characteristics. Specifically, the proposed model embeds each node with a latent vector shared among layers and a layer-specific effect for each layer; both elements together with a layer-specific connectivity matrix determine edge formations. To fit the model, we develop a projected gradient descent algorithm for efficient parameter estimation. We also establish theoretical properties of the maximum likelihood estimators and show that the upper bound of the common latent structure’s estimation error is inversely proportional to the number of layers under mild conditions. The superior performance of the proposed model is demonstrated through simulation studies and applications to two real-world data examples.

## Session 66: Bridging the gap between complex data and public health policies: methods and applications

### VC-BART: Varying Coefficients

♦ Sameer Deshpande<sup>1</sup>, Ray Bai<sup>2</sup>, Cecilia Balocchi<sup>2</sup> and Jennifer Starling<sup>3</sup>

<sup>1</sup>MIT

<sup>2</sup>University of Pennsylvania

<sup>3</sup>The University of Texas at Austin  
sameerd@alum.mit.edu

The linear varying coefficient (VC) model generalizes the conventional linear model by allowing the additive effect of each covariate  $X$  on the outcome  $Y$  to vary as a function of additional effect modifiers  $Z$ . Since its introduction, VC models have been used widely to analyze longitudinal, spatiotemporal, and economic data. While there are many existing procedures for fitting such a model when the effect modifier  $Z$  is a scalar (typically time), there has been comparatively less development for settings with multivariate  $Z$ . In this work, we present an extension of Bayesian Additive Regression Trees (BART) to the varying coefficient model for applications in which we might reasonably suspect covariate effects vary systematically with respect to interactions between multiple modifiers. We derive a straightforward Gibbs sampler that also allows for correlated residual errors. We further build on recent theoretical advances for the varying-coefficient model and BART to derive posterior concentration rates under our model. We demonstrate our method on several econometric and spatiotemporal examples. We will then discuss several extensions and potential applications of our methodology to public health data.

### Bounds for local average treatment effects in instrumental variable analyses of mobile interventions

Andrew Spieker

Vanderbilt University Medical Center  
andrew.spieker@vumc.org

Estimation of local average treatment effects requires the exclusion restriction to hold in cases where we are unwilling to rule out unmeasured confounding; namely, treatment benefit is assumed to be mediated through the post-randomization variable being conditioned upon. Recently, there has been interest in mobile health interventions to provide healthcare support; such studies can feature one-way and/or two-way content, the latter of which allowing subjects to engage with the intervention in a way that can be objectively measured (e.g., text message response rate). It is likely that the treatment effect could be explained both by receipt of the intervention content and by responding to it. When seeking to characterize average causal effects conditional on post-randomization engagement, the exclusion restriction is therefore all but surely violated. We propose a conceptually intuitive sensitivity analysis procedure for this setting that gives rise to bounds on local average treatment effects. Simulation studies reveal this approach to have good finite-sample behavior and to recover parameters under correct specification of the sensitivity parameter.

### Integrating Sample Relatedness Information into Latent Class Models: A Tree-Structured Shrinkage Approach

Zhenke Wu

University of Michigan, Ann Arbor  
zhenkewu@gmail.com

This talk is concerned with probabilistic assignment of multivariate binary observations into unobserved classes with scientific meanings. We focus on the setting where additional sample relatedness information is available and represented by an edge-weighted rooted

tree. Leaves in the given tree represent observations so that, based on the external information, more similar observations have shorter distances in the tree. We proposed a novel data integrative extension to classical latent class models (LCMs) with tree-structured shrinkage that enables 1) borrowing of information across clades, 2) data-driven grouping of observations that can be described by similar LCM parameters, and 3) probabilistic class assignment for each observation given the observed measurements. Extensive simulations show more accurate probabilistic assignment than alternatives that coarsen or ignore the additional sample relatedness information. We demonstrate the method via a *E. coli* data set where sample relatedness information is summarized by a phylogenetic tree. We discuss how the methods can extract essential information for designing more effective biocontrol strategies to reduce extraintestinal *E. coli* infections.

### Bayesian Latent Class Models for Verbal Autopsy Data from Multiple Domains.

Zehang Li

University of California, Santa Cruz  
lizhang@gmail.com

Verbal autopsy (VA) is a survey-based tool for assigning a cause to deaths when traditional autopsy and cause certification are not available. It has been routinely used for mortality surveillance in low-resource settings. In the last decade, several statistical and machine learning methods for inferring cause-of-death using VA data have been developed. Generalizability has been a common challenge with most of the probabilistic VA algorithms, as data collected from different domains, e.g., locations or time periods, often exhibit different relationships between causes and symptoms. As a result, the choice of training data has strong implications on the performance of VA algorithms. In this talk, I will present statistical approaches to characterize the joint distribution of symptoms while accounting for the heterogeneity of data from different domains. We propose a novel latent class model that classifies causes-of-death by learning the similarities between the new domain and the existing domains. I will demonstrate the performance and interpretability of the method using a gold-standard VA dataset collected from multiple study sites.

## Session 67: Advanced Bayesian methods in Biostatistics

### A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes

♦ Matthew Koslovsky<sup>1</sup> and Marina Vannucci<sup>2</sup>

<sup>1</sup>Rice University

<sup>2</sup>Rice University  
mkoslovsky@rice.edu

One of the major research questions regarding human microbiome studies is the feasibility of designing interventions that modulate the composition of the microbiome to promote health and cure disease. This requires extensive understanding of the modulating factors of the microbiome, such as dietary intake, as well as the relation between microbial composition and phenotypic outcomes, such as body mass index (BMI). Previous efforts have modeled these data separately, employing two-step approaches that can produce biased interpretations of the results. Here, we propose a Bayesian joint model that simultaneously identifies clinical covariates associated with microbial composition data and predicts a phenotypic response using information contained in the compositional data. Using spike-and-slab priors, our approach can handle high-dimensional compo-

sitional as well as clinical data. Additionally, we accommodate the compositional structure of the data via balances and overdispersion typically found in microbial samples. We apply our model to understand the relations between dietary intake, microbial samples, and BMI. In this analysis, we find numerous associations between microbial taxa and dietary factors that may lead to a microbiome that is generally more hospitable to the development of chronic diseases, such as obesity.

### Graphical Models for Data Integration and Mediation Analysis

*Min Jin Ha*

The University of Texas MD Anderson Cancer Center  
mjha@mdanderson.org

Integrative network modeling of data arising from multiple genomic platforms provides insight into the holistic picture of the interactive system, as well as the flow of information across many disease domains. The basic data structure consists of a sequence of hierarchically ordered datasets for each individual subject, which facilitates integration of diverse inputs, such as genomic, transcriptomic, and proteomic data. A primary analytical task in such contexts is to model the layered architecture of networks where the vertices can be naturally partitioned into ordered layers, dictated by multiple platforms, and exhibit both undirected and directed relationships. We propose a multi-layered Gaussian graphical model (mlGGM) to investigate conditional independence structures in such multi-level genomic networks. We use a Bayesian node-wise selection approach that coherently accounts for the multiple types of dependencies in mlGGM, that is used for finding causal factors for outcome variables via mediation analysis.

### Flexible and informative clustering of microbiome data

*Christine Peterson*

The University of Texas MD Anderson Cancer Center  
cbpeterson@mdanderson.org

The microbiome plays an important role in human health and disease. Microbiome data poses a number of statistical challenges due to its unique structure: specifically, the observed data are counts with a fixed sum constraint which can be organized into a taxonomic or phylogenetic tree structure. Unsupervised clustering is often used to identify naturally occurring groups of subjects with similar microbiome profiles. However, popular machine learning based microbiome clustering methods may fail in certain settings. I will discuss a flexible and informative Bayesian approach to clustering that takes into account the unique structure of microbiome data, and has key performance advantages over existing methods.

### Dependent Mixtures: Modeling Cell Lineage

♦ *Giorgio Paulon, Carlos Paganizani and Peter Müller*

The University of Texas at Austin  
giorgio.paulon@utexas.edu

We introduce dependent mixture models for model-based inference in mixtures when the cluster locations are naturally connected by a spanning tree. The motivating application is inference for cell lineage data on the basis of single cell RNA-seq data for cells across different levels of cell differentiation. The terms of the mixture model are interpreted as representing distinct cell types, including a known root cell population and final differentiated cells. We propose prior models based on prior shrinkage of the cumulative length of a minimum spanning tree (MST) of cluster centers.

## Session 68: Leveraging Real-World Data in Comparative Effectiveness Research

### Real world data, machine learning and causal inference

*Jie Chen*

Overland Pharma  
jiechen0713@gmail.com

There has recently been an increasing interest in applying statistical and machine learning algorithms to real-world data (RWD) for causality assessment. One of the major challenges in analyzing RWD is confounding bias that can lead to spurious association between an exposure and an outcome. Although there are several different approaches (such as propensity score matching and stratification) to adjusting for confounding bias in causal inference, this presentation will focus on structural causal model (SCM) based machine learning methodologies including target learning and reinforcement learning and application of these approaches to RWD for causal inference.

### Estimation and variable selection for conditional causal effect: a dimension reduction approach

*Zonghui Hu*

National Institutes of Health  
huzo@niaid.nih.gov

Medical studies usually involve participants representing the wide heterogeneous population. The conditional causal effect, treatment effect conditional on baseline characteristics, is of practical importance. Its estimation is subject to two challenges. First, the causal effect is not observable in any individual due to the counterfactual nature. Second, observational studies tend to involve high-dimensional baseline characteristics to satisfy the ignorable treatment selection assumption and to include all possible predictors of the outcome. As a consequence, correct model specification is nearly impossible. In this work, we address the first challenge through the pseudo-response. To cope with high dimensionality, we estimate the conditional causal effect via the “characteristic score” — a parsimonious representation of the baseline covariate influence on treatment benefit — through nonparametric dimension reduction, which is a sparse dimension reduction preceded by covariate pre-screening. The characteristic score characterizes an individual in terms of the potential benefit from a treatment. This approach is applied to an HIV study for assessing the benefit of antiretroviral regimens and identifying the beneficiary subpopulation.

### Multilevel-Multiclass Graphical Model for Correlated Network

*Inyoung Kim*

Virginia Tech  
inyoungk@vt.edu

The Gaussian graphical model has been a popular tool for investigating the conditional dependency structure between random variables by estimating sparse precision matrices. However, the ability to investigate the conditional dependency structure when a two-level structure exists among the variables is still limited. Some variables are considered as higher-level variables while others are nested in these higher-level variables - the latter are called lower-level variables. Higher-level variables are not isolated; instead, they work together to accomplish certain tasks. Therefore, our main interest is to simultaneously explore conditional dependency structures among higher-level variables and among lower-level variables. Given two-level data from heterogeneous classes, we propose a method to jointly estimate the two-level Gaussian graphical models across multiple classes, so that common structures in terms of

the two-level conditional dependency is shared during the estimation procedure, yet unique structures for each class are retained as well. We also demonstrate the advantages of our approach using breast cancer patient data.

### Adjusting for Population Differences Using Machine Learning Methods

Zhiwei Zhang

National Institutes of Health  
zhiwei.zhang@nih.gov

The use of real-world data for medical treatment evaluation frequently requires adjusting for population differences. We consider this problem in the context of estimating mean outcomes and treatment differences in a well-defined target population, using clinical data from a study population that overlaps with but differs from the target population in terms of patient characteristics. The current literature on this subject includes a variety of statistical methods, which generally require correct specification of at least one parametric regression model. In this work, we propose to use machine learning methods to estimate nuisance functions and incorporate the machine learning estimates into existing doubly robust estimators. This leads to nonparametric estimators that are root-n consistent, asymptotically normal, and asymptotically efficient under general conditions. Simulation results demonstrate that the proposed methods perform well in realistic settings. The methods are illustrated with a cardiology example concerning aortic stenosis.

### Session 70: Design and Statistical Issues for Pediatric Oncology Trials

#### PA-CRM: A Continuous Reassessment Method for Pediatric Phase I Trials with Concurrent Adult Trials

Yimei Li<sup>1</sup> and Ying Yuan<sup>2</sup>

<sup>1</sup>University of Pennsylvania

<sup>2</sup>The University of Texas MD Anderson Cancer Center  
yimeili@penmedicine.upenn.edu

Pediatric phase I trials are usually carried out after the adult trial has started, but not completed yet. As the pediatric trial progresses, in light of the accrued interim data from the concurrent adult trial, the pediatric protocol often is amended to modify the original pediatric dose escalation design. This frequently is done in an ad hoc way, interrupting patient accrual and slowing down the trial. We develop a pediatric continuous reassessment method (PA-CRM) to streamline this process, providing a more efficient and rigorous method to find the MTD for pediatric phase I trials. We use a discounted joint likelihood of the adult and pediatric data, with a discount parameter controlling information borrowing between pediatric and adult trials. According to the interim adult and pediatric data, the discount parameter is adaptively updated using the Bayesian model averaging method. We examine the PA-CRM through simulations, and compare it with the two alternative approaches, which ignore adult data completely or simply pool it together with the pediatric data. The results demonstrate that the PA-CRM has good operating characteristics and is robust to various assumptions.

#### A Review of the experience of pediatric written requests issued for oncology drug products

Jingjing Ye

BeiGene  
jingjing.ye@beigene.com

There are well recognized challenges to pediatric oncology drug development. We will present and evaluate the experience with writ-

ten requests issued by FDA as authorized by the Best Pharmaceuticals for Children Act (BPCA) as to the current status, successful completion, study designs, statistical analysis plans, and evidentiary contribution to product approval or labeling changes. The results are based on 44 written requests that were issued by the FDA for pediatric clinical trials of new or recently approved cancer drugs which were completed between 2001 to 2018. We describe the types of studies in these written requests and the patient populations enrolled as well as details on study design, endpoints, investigational drug formulation and plans for pediatric-appropriate formulations, extrapolation considerations, statistical analysis plan, evidentiary standards for efficacy and safety determination. We will also review potential challenges in pediatric clinical studies and future trial designs considerations.

#### BOIN12: Bayesian Optimal Interval Phase I/II Trial Design for Utility-Based Dose Finding in Immunotherapy and Targeted Therapies

Ruitao Lin<sup>1</sup>, Yahong Zhou<sup>1</sup>, Dianel Li<sup>2</sup>, Fangrong Yan<sup>3</sup> and Ying Yuan<sup>1</sup>

<sup>1</sup>The University of Texas MD Anderson Cancer Center

<sup>2</sup>Bristol-Myers Squibb

<sup>3</sup>China Pharmaceutical University  
yyuan@mdanderson.org

For immunotherapy such as checkpoint inhibitors and CAR-T cell therapy, as the efficacy does not necessarily increase with the dose, the maximum tolerated dose (MTD) may not be the optimal dose for treating patients. For these novel therapies, the objective of dose-finding trials is to identify the optimal biological dose (OBD) that optimizes patients' risk-benefit tradeoff. We propose a simple and flexible Bayesian optimal interval phase I/II (BOIN12) trial design to find the OBD that optimizes the risk-benefit tradeoff. The BOIN12 design makes the decision of dose escalation and de-escalation by simultaneously taking account of efficacy and toxicity, and adaptively allocates patients to the dose that optimizes the toxicity-efficacy tradeoff. Compared to existing phase I/II dose-finding designs, the BOIN12 design is simpler to implement, has higher accuracy to identify the OBD, and allocates more patients to the OBD. One of the most appealing features of the BOIN12 design is that its adaptation rule can be pre-tabulated and included in the protocol. During the trial conduct, clinicians can simply look up the decision table to allocate patients to a dose without complicated computation. User-friendly software is freely available at [www.trialdesign.org](http://www.trialdesign.org) to facilitate the application of the BOIN12 design.

### Session 71: Enhancing RCT using Real World Evidence

#### Creating A Synthetic Control Arm Using Propensity Score Analysis

Zailong Wang, Zhuqing Yu and Lanju Zhang

ABBVIE

zailong.wang@yahoo.com

Propensity score matching is a statistical technique to improve the accuracy of treatment effect estimates by matching on relevant covariates. In a single-arm clinical study without a control group, propensity score analysis could be used to create a synthetic control arm from either historical trial data or real-world data. Hence the treatment effect could be estimated between the treatment group in the study and the synthetic control group. In this presentation, we will introduce propensity score methods and the motivation for creating a synthetic control arm, followed by real data analyses

to illustrate the applications and compare the differences between these methods. We will conclude with our recommendations based on our experiences of such applications.

### Use of Hybrid Control Arms with Adaptive Power Priors for Time-to-Event Data

Matthew Psioda

UNC-CH

matt\_psioda@unc.edu

We consider an extension of the power prior where the borrowing parameter is determined separately for each external control based on a compatibility metric derived using internal control data and the data for the external control. Compatibility is assessed using the predictive distribution for each external control's outcome determined by integrating over the outcome model parameter with respect to the posterior distribution resulting from analysis of the internal control data. Case-specific borrowing parameters allow for differential borrowing across the set of external controls and can be useful in cases where some external control data may be contaminated, e.g. due to erroneous data in electronic health records. We develop the approach with an eye on application to time-to-event regression models for outcomes subject to right censoring. This scenario is particularly challenging due to the outcome of interest not always being observed making compatibility as described above more challenging to characterize.

### Analytic framework for non-randomized single-arm clinical trials with external RWD control

♦ Hongwei Wang, Yixin Fang and Weili He

AbbVie

hongwei.wang@abbvie.com

Real-world data (RWD) is playing an increasingly important role in drug development from early in discovery throughout the life-cycle management. This includes improving the efficiency of clinical trial design and conduct. In many scenarios, a concurrent control arm may not be viable for ethical or practical consideration, and inclusion of an external control arm can greatly facilitate the decision-making and interpretation of findings. To address the inherent confounding due to lack of randomization, propensity-score matching method has the advantages of separating the design from analysis and providing the ability to explicitly examine the degree of overlap in confounders. Within the framework of causal inference, many alternatives have been proposed with desirable theoretical properties. In this talk, we focus on inverse probability of treatment weighted (IPTW), augmented IPTW, G-formula, targeted MLE (TMLE), and TMLE coupled with super learner. Their performances in terms of bias reduction and statistical precision are assessed in a simulation study including scenarios when underlying assumptions are violated or models are mis-specified. Practical considerations are given for their implementation.

### Practical Considerations in Design and Execution of RWE Studies

Yijie Zhou

Vertex Pharmaceuticals

yijie.zhou@vrtx.com

Undoubtedly there has been increasing interest in evidence generation from real world studies to supplement clinical trial data. In parallel to the statistical methodology development in this area, the unique operational features of RWE studies require different practical considerations in study design, data collection and analysis approaches. In this talk we will discuss the practical considerations in setting statistical assumptions for RWE study design, data collec-

tion deployment to enable sufficient information, and analysis plan development to accommodate the data limitation.

### Session 72: Real World Evidence for Value-Added Patient-Centric Healthcare

#### National Health and Wellness Survey exploratory cluster analysis of males 40-70 years old focused on erectile dysfunction and associated risk factors across the USA, Italy, Brazil and China

Irwin Goldstein<sup>1</sup>, Amir Goren<sup>2</sup>, Ryan Liebert<sup>2</sup>, Wing Yu Tang<sup>3</sup> and ♦Tarek Hassan<sup>4</sup>

<sup>1</sup>Alvarado Hospital

<sup>2</sup>Kantar

<sup>3</sup>Pfizer

<sup>4</sup>Viatrix

tarek.hassan@viatrix.com

**OBJECTIVES:** Prior studies of erectile dysfunction (ED) tend to narrowly focus on relationships with specific comorbidities, rather than evaluating a more comprehensive array of risk factors and assessing naturalistic patterns among them. This study identifies natural clusters of male characteristics from a general population sample per country, quantifies ED dynamics in these profiles and compares profiles across the US, Italy, Brazil and China samples. **METHODS:** National Health and Wellness Survey 2015 and 2016 patient-reported data on men aged 40-70 years (USA n = 15,652; Italy n = 2,521; Brazil n = 2,822; China n = 5,553) were analysed. Hierarchical agglomerative clustering identified clusters where predictors included demographics, health characteristics/behaviours, ED risk factors and provider visits in the past 6 months. Multinomial logistic regression assessed the independent utility of variables in predicting cluster membership, compared with the healthiest control cluster per country. **RESULTS:** Different natural clusters were found across countries, with four clusters for the USA, Italy and China and three clusters for Brazil. Age, income, employment, health behaviours and ED risk factors predicted different cluster membership across countries. In the USA, Italy and Brazil, younger clusters were predicted by ED, unhealthy behaviours and ED risk factors. Unique cluster profiles were identified in China, with ED and ED risk factors (aside from hypertension) not predicting cluster membership, while socio-demographics and health behaviours were strongly predictive. **CONCLUSIONS:** Natural cluster profiles revealed notable ED rates among adult males of age 40-70 in four different countries. Clusters were mainly predicted by unhealthy behaviours, ED risk factors and ED, regardless of level or presence of positive health characteristics and behaviours. This analysis identified meaningful subgroups of men with heightened ED risk factors, which can help healthcare providers to better recognise specific populations with the greatest need for intervention.

#### Evidence and Access with Uncertainty: Establishing Value with Evolving Evidence

♦ Joseph Cook<sup>1</sup> and Adam Heathfield<sup>2</sup>

<sup>1</sup>Viatrix

<sup>2</sup>Pfizer

joseph.cook@viatrix.com

**Abstract:** For new treatments, there is a serious dilemma of evidence vs access. This dilemma revolves around the balance between demanding additional evidence of value, which will delay patient access, or allowing patients access now with the current evidence and more uncertainties. The clinical, economic, and ethical aspects of



the dilemma have raised questions with implications both for individual products and the longer-term view of overall innovation in medicines. These competing considerations further call into question the sustainability of current drug development and timely access to innovative medicines for patients. Of course, all decisions have some measure of risk and uncertainty, and the goal is never to eliminate them, but to efficiently manage them. Our focus here is on leveraging the full range of evidence available to support access decisions. Clinical models and economic models associated with the new treatment will all incorporate some measure of variability in observed outcomes. We can think of this variability as Knightian risk that is generally quantifiable in the context of the models. The risk that the cost of the treatment exceeds its average benefit in the population can, trivially, be reduced with a lower cost for the treatment, but this has the concomitant effect of reducing the likelihood such treatments are developed for the market. The average effect and associated value seem the most equitable and balanced approach to the quantifiable risk, but more is needed for the unquantifiable uncertainties associated with generalizing the results of the randomized clinical trials to the populations at large. Adaptive access and evolving evidence can provide the solution, by allowing responsiveness in access and reimbursement to the best available evidence over time and separating the concerns associated with risk and uncertainty.

#### **Real World Evidence on the Impact of Sertraline Daily Treatment Regimen on Medication Adherence and Persistence in Patients with Major Depressive Disorder or Obsessive-Compulsive Disorder**

*Gang Wang<sup>1</sup>, Tianmei Si<sup>2</sup>, ♦Joseph Imperato<sup>3</sup>, Li Li Yang<sup>3</sup>, Kelly Zou<sup>3</sup>, Ying Jin<sup>3</sup>, Elizabeth Pappadopulos<sup>3</sup>, Lei Yan<sup>3</sup> and Wei Yu<sup>3</sup>*

<sup>1</sup>Beijing Anding Hospital Affiliated to Capital Medical University

<sup>2</sup>Peking University

<sup>3</sup>Viatris

joseph.imperato@viatris.com

Nonadherence and poor persistence with antidepressant medications can negatively impact treatment outcomes in patients with major depressive disorder (MDD) or obsessive-compulsive disorder (OCD). Sertraline is a first-line treatment for MDD and OCD with a well-established safety and efficacy profile. However, limited data are available on the influence of daily treatment regimen on adherence with sertraline. This real-world data (RWD) analysis explores the impact of sertraline dose stabilization and number of switches prior to dose stabilization, daily pill quantity (pill burden), and other patient-level variables on adherence/persistence and disease-specific healthcare utilization outcomes. A retrospective RWD analysis was carried out using the IQVIA PharMetrics Plus claims database, comprising commercially insured patients in the US. The index therapy date was when sertraline dose became stabilized, defined as >90 days on same dose after the initial prescription during the selection window from 4/1/2013 to 3/31/2019. Eligible patients had an International Classification of Diseases, Ninth or Tenth Revision, Clinical Modification (ICD-9-CM or ICD-10-CM) coding for MDD or OCD and eligible patients were analyzed in 5 cohorts: 1×50-mg pill, 1×100-mg pill, 2×50-mg pills, 3×50 mg pills, and 1.5×100-mg pills. Baseline variables captured were demographics, comorbidities, healthcare utilization and other clinical characteristics. Outcomes were adherence (calculated using medication-possession ratio [MPR]) and persistence to treatment, and disease-specific healthcare usage. Descriptive statistics and multivariate regressions were used. Real world evidence provides an understanding of the potential impacts on nonadherence and poor

persistence due to some characteristics of the treatment regimen, which may help improve clinical management of p

#### **Harnessing real-world evidence to reduce the burden of non-communicable disease: health information technology and innovation to generate insights**

*Kelly Zou<sup>1</sup>, ♦Jim Li<sup>1</sup>, Lobna A.salem<sup>1</sup>, Joseph Imperato<sup>1</sup>, Jon Edwards<sup>2</sup> and Amrit Ray<sup>3</sup>*

<sup>1</sup>Viatris

<sup>2</sup>Envision Pharma Group

<sup>3</sup>Pfizer

jim.li@viatris.com

Noncommunicable diseases (NCDs) are the leading causes of mortality and morbidity across the world and factors influencing global poverty and slowing economic development. We summarize how the potential power of real-world data (RWD) and real-world evidence (RWE) can be harnessed to help address the disease burden of NCDs at global, national, regional and local levels. RWE is essential to understand the epidemiology of NCDs, quantify NCD burdens, assist with the early detection of vulnerable populations at high risk of NCDs by identifying the most influential risk factors, and evaluate the effectiveness and cost-benefits of treatments, programs, and public policies for NCDs. To realize the potential power of RWD and RWE, challenges related to data integration, access, interoperability, standardization of analytical methods, quality control, security, privacy protection, and ethical standards for data use must be addressed. Finally, partnerships between academic centers, governments, pharmaceutical companies, and other stakeholders aimed at improving the utilization of RWE can have a substantial beneficial impact in preventing and managing NCDs. (Keywords: real-world data; real-world evidence; population health; health information technology; noncommunicable disease; disease burden; risk factors; data science.)

#### **Session 73: High-dimensional statistical learning in big-data of human genetics**

##### **Gene Network Analysis with Single Cell RNA Sequencing Data**

*♦Fei Zou and Meichen Dong*

UNC-CH

feizou@email.unc.edu

Cutting edge single-cell sequencing data enables researchers to study networks and pathways for complex tissues and under different conditions. Current existing network analysis approaches treat single cells equally distant from each other without exploiting the inter-cell relationships. Here, we propose a framework to construct gene networks based on fused-lasso regression for ordered single cells to construct more robust and efficient networks. The proposed method is first illustrated on synthetically simulated data and then applied to a medulloblastoma scRNA-seq dataset.

##### **Multi-omics data integration with kernel fusion**

*Haitao Yang<sup>1</sup> and ♦Yuehua Cui<sup>2</sup>*

<sup>1</sup>Hebei Medical University

<sup>2</sup>Michigan State University

cuiy@msu.edu

High throughput omics data are generated almost with no limit nowadays. It becomes increasingly important to integrate different omics data types to disentangle the molecular machinery of complex diseases with the hope for better disease prevention and treatment. In this talk, I will briefly introduce the idea of kernel fusion for data

integration. We focus on a fused kernel partial least squares (fK-PLS) model for disease classification with multi-level omics data. The fused kernel can deal with effect heterogeneity in which different omics data types may have different effect contribution to the trait of interest. We optimize the kernel parameters and kernel weights with the genetic algorithm (GA). The proposed GA-fKPLS model can substantially improve disease classification performance by integrating multiple omics data types, demonstrated via simulation studies and real data analysis.

#### **MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations**

♦ *Zifan Zhu, Jie Ren, Sonia Michail and Fengzhu Sun*

University of Southern California, Los Angeles  
zifanzhu@usc.edu

We develop a metagenomic data analysis pipeline, MicroPro, that takes into account all reads from known and unknown microbial organisms and associates viruses with complex diseases. We utilize MicroPro to analyze four metagenomic datasets relating to colorectal cancer, type 2 diabetes, and liver cirrhosis and show that including reads from unknown organisms significantly increases the prediction accuracy of the disease status for three of the four datasets. We identify new microbial organisms associated with these diseases and show viruses play important prediction roles in colorectal cancer and liver cirrhosis, but not in type 2 diabetes. MicroPro is freely available at <https://github.com/zifanzhu/MicroPro>.

#### **Joint modeling of bacterial and fungal network in Alcoholic hepatitis**

*Xinlian Zhang*

University of California San Diego  
xizhang@ucsd.edu

Alcoholic hepatitis is one of the most severe alcohol-related liver diseases and is generally associated with high mortality. Changes in the intestinal microbiota composition, which is often due to chronic alcohol usage causes, is known to contribute to the development and progression of alcohol-related liver disease. In this project, we studied the intestinal microbiota in a cohort of patients with alcoholic hepatitis, patients with alcohol use disorder, and nonalcoholic controls and analyze the gut bacterial and fungal sequence data of the fecal samples of the patients. Using these data, we discuss and explore the joint modeling of the bacterial and fungal networks. In the future, we intend to include the virome data of these patient samples and further extend the analysis.

### **Session 74: Precision oncology trials: challenges and opportunities**

#### **On Design and Analysis of Biomarker-Integrated Clinical Trials with Adaptive Threshold Detection and Patient Enrichment**

♦ *Xiaofei Wang<sup>1</sup>, Ting Wang<sup>2</sup>, Stephen George<sup>1</sup> and Haibo Zhou<sup>2</sup>*

<sup>1</sup>Duke University

<sup>2</sup>Univ. of North Carolina at Chapel Hill  
xiaofei.wang@duke.edu

In biomarker-integrated clinical trials, an optimal biomarker threshold is often not available at patient randomization. Several adaptive threshold-detecting designs have been proposed, in which the target population is adaptively learned and the enrollment criteria designed by the cutoff of a continuous biomarker is adaptively updated. These existing designs have largely ignored investigation of the estimation accuracy of the biomarker threshold and the enrollment criteria, and the estimation of treatment effects. In this pa-

per, we propose a new adaptive threshold detection and enrichment design, in which the optimal true biomarker threshold is regularly estimated and updated, and the enrichment of the patients who are benefiting from the experiment therapy is optimized in a trade-off between the size of the positive population and the magnitude of the treatment effect of that population. Early termination for futility is also allowed based on the predictive probability of success for biomarker-positive patients. Valid test and estimation on treatment effects overall or in patient subgroups are also studied. Simulation results demonstrate that the proposed design has several advantages compared to existing designs, including more accurate estimation of the biomarker threshold and significant reduction in cost and time. The proposed design is illustrated with a lung cancer study trial with PD-L1 as a potential biomarker for the efficacy of immunotherapy.

#### **Group sequential enrichment designs based on adaptive regression of response and survival time on high dimensional covariates**

♦ *Yeonhee Park<sup>1</sup>, Suyu Liu<sup>2</sup>, Peter Thall<sup>2</sup> and Ying Yuan<sup>2</sup>*

<sup>1</sup>Medical University of South Carolina

<sup>2</sup>The University of Texas MD Anderson Cancer Center  
ypark56@wisc.edu

Precision medicine relies on the idea that only a subpopulation of patients are sensitive to a targeted agent and thus may benefit from it. In practice, based on pre-clinical data, it often is assumed that the sensitive subpopulation is known and the agent is substantively efficacious in that subpopulation. Subsequent patient data, however, often show that one or both of these assumptions are false. This paper provides a Bayesian randomized group sequential enrichment design to compare an experimental treatment to a control based on survival time. Early response is used as an ancillary outcome to assist with adaptive variable selection, enrichment, and futility stopping. The design starts by enrolling patients under broad eligibility criteria. At each interim decision, submodels for regression of response and survival time on a possibly high dimensional covariate vector and treatment are fit, variable selection is used to identify a covariate subvector that characterizes treatment-sensitive patients and determines a personalized benefit index, and comparative superiority and futility decisions are made. Enrollment of each cohort is restricted to the most recent adaptively identified treatment-sensitive patients. Group sequential decision cutoffs are calibrated to control overall type I error and account for the adaptive enrollment restriction. The design provides an empirical basis for precision medicine by identifying a treatment-sensitive subpopulation, if it exists, and determining whether the experimental treatment is substantively superior to the control in that subpopulation. A simulation study shows that the proposed design accurately identifies a sensitive subpopulation if it exists, yields much higher power than a conventional group sequential design, and is robust.

#### **Bayesian Semi-parametric Design (BSD) for Adaptive Dose-finding with Multiple Strata**

♦ *Rachael Liu<sup>1</sup>, Jianchang Lin<sup>1</sup>, Mo Li<sup>2</sup>, Veronica Bunn<sup>1</sup> and Hongyu Zhao<sup>2</sup>*

<sup>1</sup>Takeda Pharmaceuticals

<sup>2</sup>Yale University  
yue.liu@takeda.com

In the era of precision medicine, it is of increasing interest to consider multiple strata (e.g. indications, regions or subgroups) within a single oncology dose-finding study when identifying the maximum tolerated dose (MTD). We propose two Bayesian semi-parametric models (BSD) for dose-finding with multiple strata to allow for

both adaptively dosing patients based on various toxicity profiles and efficient identification of the MTD for each stratum. We develop non-parametric priors based on the Dirichlet process to allow for a flexible prior distribution and negate the need for a pre-specified exchangeability parameter. The two BSD models are built under differing prior beliefs of strata heterogeneity and allow for appropriate borrowing of information across similar strata. Simulation studies are performed to evaluate the BSD model performance by comparing with existing methods, including the fully stratified, exchangeability, and exchangeability-non-exchangeability models. In general, our BSD models outperform the competing methods in correctly identifying the MTD for different strata and necessitate a smaller sample size to determine the MTD. The BSD models are robust to various heterogeneity assumptions and can be easily extended to other binary and time to event endpoints.

## Session 75: Statistical Advancement and Challenges in Cell Therapy Development

### Statistical Challenges In CAR-T Cell Therapy Development

*Daniel Li*

BMS

daniel.li@junotherapeutics.com

In the past few years, promising antitumor activity has been seen with cellular therapies. The most striking effect has been with CD19-targeted chimeric antigen receptor (CAR) T cells in relapsed or refractory B-cell malignancies. This new modality of treatment has recently been validated by the FDA with recent approvals of two CAR T-cell therapies (Yescarta and Kymriah). Unlike small molecules or biologics, cellular therapies are highly personalized products with unique features. These features pose additional statistical challenges across all phases of development for a cellular therapy. Specifically, challenges that will be discussed will include 1) phase 1 dose-finding and whether the standard dose finding approach developed for other targeted therapies are suitable for CAR T-cell products and other cellular therapies; 2) phase 2 and randomized phase 3 studies and how to handle manufacturing failures, timing of randomization, non-proportional hazards, and delayed treatment effects; 3) how to utilize artificial intelligence and machine learning to manage the large amount of clinical, translation and manufacturing, and product data to improve the manufacturing process or predict patients who will likely experience benefit or toxicity; and 4) discuss the statistical challenges that arise with the evolving manufacturing processes that occur in the development of cellular therapies.

### Statistical review of gene and cell therapies and related research topics

*Xue (Mary) Lin*

The Food and Drug Administration

xue.lin@fda.hhs.gov

In this presentation, first we will talk about the statistical issues in the biological licensure application review of KYMRIAH®, which was the first CAR-T therapy approved by the FDA. In addition, we will discuss statistical issues related to the timing of randomization in the study design of gene and cell therapy trials, all based on INDs we have reviewed. At the end, we will touch upon some research topics we are working on.

### Interpretation of the Treatment Effect in the Context of Complex Treatment Strategy and Methodological considerations for

### CAR-T clinical trials

*YiYun (Michael) Zhang*

Autolus Ltd

m.zhang@autolus.com

Autologous chimeric antigen receptor (CAR) T cells therapies is changing the therapeutic landscape in haematological malignancies such as acute lymphoblastic leukemia (ALL) and non-Hodgkin's lymphoma (NHL). Traditionally, treatment strategies of these disease involve complex combinations of chemo and/or immunotherapies. Depending on the disease setting, stem cell transplant (SCT) is considered, and is regarded as a potential curative option. However, these existing therapies can be associated with high rates of treatment related mortality and potential long term clinical complications. Therefore it is of great interest to evaluate the effect of CAR T treatment either as a stand-alone therapy, or as part of a treatment strategy that includes subsequent consolidation (e.g. by SCT) therapy. The disease setting and target treatment effect of interest must be taken into account in the design and analysis of CAR-T clinical trials. The author will discuss these methodological considerations under the estimand framework (ICH E9 addendum).

### Statistics Considerations to Find Optimal Dose of CAR-T Cell Therapy

*Xiaoling Wu*

Legend Biotech

xiaoling.wu@legendbiotech.com

Phase I oncology trials usually test an experimental drug at a fixed sequence of dose level, and aim to identify the maximum tolerated dose (MTD). The commonly used dose-escalation method falls into two classes, rule-based and model based. These methods consider the dose-limiting toxicity data only and implicitly assume a monotone dose-response relationship. But these assumption and concept may not entirely apply to cell therapy. It is therapy in which cellular material is injected into a patient; For example, T cells capable of fighting cancer cells via cell-mediated immunity may be injected in the course of immunotherapy. Under such setting, patient will receive a live product of intact cells manufactured in a lab. And they might respond rather quickly after infusion. A joint model to determine an optimal dose for next step of clinical development may come appropriate and handy. This work will introduce a simple method of incorporating toxicity and efficacy data to determine the optimal dose in a mixed patient population. A Bayesian logistic regression will be used to quantify the dose-toxicity and dose-efficacy relations. A utility score will then be computed for each dose using the dose limiting toxicity (DLT) and response status. And to further guide the determination of dose for next step of clinical development. A few simulated case studies will be conducted with various toxicity and efficacy profiles. Performance in terms of accuracy and safety will be assessed.

## Session 76: Genetics and Genomics: Methodology and Applications

### Ordered Multinomial Regression for Genetic Association Analysis of Ordinal Phenotypes at Biobank

*Jin Zhou*

University of Arizona

jzhou@email.arizona.edu

Logistic regression is the primary analysis tool for binary traits in genome-wide association studies (GWAS). Multinomial regression extends logistic regression to multiple categories. However, many phenotypes more naturally take ordered, discrete values. Examples

include (1) subtypes defined from multiple sources of clinical information and (2) derived phenotypes generated by specific phenotyping algorithms for electronic health records (EHR). GWAS of ordinal traits have been problematic. Dichotomizing can lead to a range of arbitrary cut off values, generating inconsistent, hard to interpret results. Using multinomial regression ignores trait value hierarchy and potentially loses power. Treating ordinal data as quantitative can lead to misleading inference. To address these issues, we analyze ordinal traits with an ordered, multinomial model. This approach increases power and leads to more interpretable results. We derive efficient algorithms for computing test statistics, making ordinal trait GWAS computationally practical for biobank scale data. Our method is available as a Julia package `OrdinalGWAS.jl`. Application to a COPDGene study confirms previously found signals based on binary case-control status, but with more significance. Additionally, we demonstrate the capability of our package to run on UK Biobank data by analyzing hypertension as an ordinal trait.

#### Gene-based association analysis of survival traits via functional regression-based mixed effect Cox models for related samples

♦ *Chi-Yang Chiu<sup>1</sup> and Ruzong Fan<sup>2</sup>*

<sup>1</sup>UTHSC

<sup>2</sup>Georgetown University  
chiu@uthsc.edu

The importance to integrate survival analysis into genetics and genomics is widely recognized, but only a small number of statisticians have produced relevant work toward this study direction. For unrelated population data, functional regression (FR) models have been developed to test for association between a quantitative/dichotomous/survival trait and genetic variants in a gene region. In major gene association analysis, these models have higher power than sequence kernel association tests. In this paper, we extend this approach to analyze censored traits for family data or related samples using FR based mixed effect Cox models (FamCoxME). The FamCoxME model effect of major gene as fixed mean via functional data analysis techniques, the local gene or polygene variations or both as random, and the correlation of pedigree members by kinship coefficients or genetic relationship matrix or both. The association between the censored trait and the major gene is tested by likelihood ratio tests (FamCoxME FR LRT). Simulation results indicate that the LRT control the type I error rates accurately/conservatively and have good power levels when both local gene or polygene variations are modeled. The proposed methods were applied to analyze a breast cancer data set from the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA). The FamCoxME provides a new tool for gene-based analysis of family-based studies or related samples.

#### Longitudinal Variant-Set Retrospective Association Test

♦ *Weimiao Wu and Zuoheng Wang*

Yale School of Public Health  
weimiao.wu@yale.edu

Set-based tests have become popular for the identification of rare genetic variants that are associated with disease traits. Burden test and variance component test are two widely used set-level tests for single time measurement. Longitudinal repeated measures have been increasingly used in genome-wide association studies. The repeated measures provide an opportunity to study the temporal development of traits and also increase the statistical power in association tests. Most of the existing variants-set association tests are based on a population model in which ascertainment is ignored. Prospective

inference with longitudinal traits and rare variants can have inflated type I error when the trait model is misspecified. Here, we propose LSRAT (Longitudinal variant-Set Retrospective Association Tests) and RSMMAT (Retrospective variant-Set Mixed Model Association Tests), two groups of retrospective variant-set tests that are constructed based on the genotype model given the phenotype and covariates. RSMMAT can be viewed as a retrospective version of the recently proposed variant-set mixed model association tests (SMMAT) and the LSRAT tests are derived under the generalized estimation equation framework. These tests have several advantages: (1) they are robust against trait model misspecification; (2) they are able to adjust both static and time-varying covariates; (3) they allow for related subjects and account for population structure; and (4) they are computationally more efficient than existing prospective approaches. Simulation studies showed that our proposed tests are robust to the trait model misspecification and gain power compared to SMMAT. We illustrated our method in the Veterans Aging Cohort Study to evaluate the association of repeated measures of alcohol use disorder with rare variants.

#### Gene-based pleiotropic analysis of multiple survival traits via functional regressions with applications to eye diseases

*Bingsong Zhang*

Georgetown University  
bz117@georgetown.edu

To analyze multiple correlated survival traits, we develop multivariate mixed effect Cox proportional hazard models by functional regressions to perform joint association analysis. The mixed effect Cox models extend fixed effect Cox models of univariate survival traits by incorporating variations and correlation of multivariate survival traits into the models. The methods are tested and refined by extensive simulation studies. We use these models to analyze bivariate traits of left and right eyes in the AMD progression data.

#### Session 77: Applications of advanced statistics and artificial intelligence to genomics and precision medicine

##### Recurrent Neural Reinforcement Learning for Counterfactual Evaluation of Public Health Interventions on the Spread of Covid-19 in the world

♦ *Qiyang Ge<sup>1</sup>, Zixin Hu<sup>1</sup>, Kai Zhang<sup>2</sup>, Tao Xu<sup>2</sup>, Shudi Li<sup>2</sup>, Wei Lin<sup>1</sup>, Li Jin<sup>1</sup> and Momiao Xiong<sup>2</sup>*

<sup>1</sup>Fudan University

<sup>2</sup>The University of Texas Health Science Center at Houston  
golden1993@icloud.com

As the Covid-19 pandemic soars around the world, there is urgent need to forecast the expected number of cases worldwide and the length of the pandemic before receding and implement public health interventions for significantly stopping the spread of Covid-19. Widely used statistical and computer methods for modeling and forecasting the trajectory of Covid-19 are epidemiological models. Although these epidemiological models are useful for estimating the dynamics of transmission of epidemics, their prediction accuracies are quite low. Alternative to the epidemiological models, the reinforcement learning (RL) and causal inference emerge as a powerful tool to select optimal interventions for worldwide containment of Covid-19. Therefore, we formulated real-time forecasting and evaluation of multiple public health intervention problems into off-policy evaluation (OPE) and counterfactual outcome forecasting problems and integrated RL and recurrent neural network (RNN) for exploring public health intervention strategies to slow down the

spread of Covid-19 worldwide, given the historical data that may have been generated by different public health intervention policies. We applied the developed methods to real data collected from January 22, 2020 to June 28, 2020 for real-time forecasting the confirmed cases of Covid-19 across the world. We forecasted that the number of laboratory confirmed cumulative cases of Covid-19 will pass 26 million as of August 14, 2020.

### Combining Artificial Intelligence and Epidemiological Models for Prediction of COVID-19 In the US

♦Tao Xu<sup>1</sup>, Zhouxuan Li<sup>1</sup>, Kai Zhang<sup>1</sup>, Hongwen Deng<sup>2</sup>, Eric Boerwinkle<sup>1</sup> and Momiao Xiong<sup>1</sup>

<sup>1</sup>The University of Texas Health Science Center at Houston

<sup>2</sup>Tulane University  
tao.xu@uth.tmc.edu

As of October, 2020, the number of cumulative cases of COVID-19 in the US exceeded 7,860,281 and included 215,955 deaths, thus causing a serious public health crisis. Curbing the spread of Covid-19 is still urgently needed. Given the lack of potential vaccines and effective medications, non-pharmaceutical interventions are the major option to curtail the spread of COVID-19. An accurate estimate of the number of cases and number of deaths from COVID-19 is crucial for planning the most effective interventions to curb the spread of COVID-19 and to reduce the deaths. Although these epidemiological models are useful for estimating the dynamics of transmission of epidemics, their prediction accuracies are quite low. Alternative to the epidemiological models, artificial intelligence (AI) emerges as a powerful tool to forecast the number of new cases and deaths from COVID-19 and select optimal interventions for containment of Covid-19 in the US. However, the results from AI are lack of interpretations. To utilize the merits of both epidemiological models and AI and overcome their limitations, we propose a new method that combine the epidemiological models and AI for forecasting the dynamics of COVID-19 in the US, calculating the time-varying number of production R and evaluating the impact of various non-pharmaceutical interventions on the curbing the spread of COVID-19 in the US. The proposed method was applied to the surveillance data of lab-confirmed Covid-19 cases in the US, University of Maryland Data (UMD) data, and Google mobility data from March 5, 2020 to October 14, 2020 in order to evaluate the contributions of social-biological factors, the Google mobility indexes, and the rate of the virus test to the number of the new cases and number of deaths from COVID-19. Our results showed that the majority of non-pharmaceutical interventions had a large effect on slowing the transmission and

## Session 78: Innovative adaptive clinical trial designs

### An optimal hybrid approach to calculate the conditional power

Jian Zhu

Servier Pharmaceuticals  
jian.zhu@servier.com

Conditional power (CP) is an important element in adaptive designs, and its performance depends on the treatment effect assumption for the remainder of the trial. To combine the strengths from two most commonly used existing methods, we propose a hybrid approach to calculate CP assuming a weighted combination of the current trend and design assumption, where the optimal weight is obtained by minimizing an expected penalty function. The penalty function can be customized to align with study-specific objectives. We demonstrated that this intuitive approach can greatly improve the precision in estimating CP. Using sample size re-estimation as examples, we

also demonstrated that the improvement in CP estimation by the proposed method can robustly minimize the consequences caused by erroneous interim decision making. Given its simplicity and desirable performance, we recommend this novel method over the existing methods.

### Innovative Adaptive Designs for Investigational Drug Development in Small Populations: A Discussion of Case Studies

♦Junjing Lin<sup>1</sup>, Godwin Yung<sup>1</sup> and Margaret Gamalo-Siebers<sup>2</sup>

<sup>1</sup>Takeda Pharmaceuticals

<sup>2</sup>Eli Lilly  
junjing.lin@takeda.com

In areas of unmet need, conducting fully powered randomized controlled trials (RCTs) may be infeasible due to enrollment challenges, timeline constraints, cost, or sometimes ethical concerns. As part of an effort to modernize approaches for drug development, in November 2019, FDA finalized the guidance for adaptive designs. This guidance emphasizes four principles for designing, conducting, and reporting results from adaptive clinical trials. It also provides advice on what information should be submitted for Bayesian and other complex trial designs. Moreover, FDA is currently conducting a 5-year Complex Innovative Trial Design (CID) Pilot Meeting Program to support the goal of facilitating and advancing the use of complex clinical trials through June 2022. With the advancement of technology, information available from sources of real-world data (RWD) have grown rapidly. Under the framework of adaptive clinical trial designs, such information can be utilized to generate evidence for making efficacy claim during interim or final analysis. Lab and genetic tests with quick turnaround can be employed to personalize treatment for patients in ongoing studies. Especially for drug development targeting small/vulnerable populations (e.g., rare diseases or children), unique challenges such as small patient numbers, heterogeneity in disease presentation, and a lack of understanding of the mechanism, call for more innovative thinking in the design and analysis of trials. For instance, how can the variability from historical data and the current trial data be accounted for during an interim analysis for different data types? How can external data from heterogeneous populations be incorporated? Are there different considerations when it comes to single-arm or multiple-arm trials? In this presentation, a few examples of rare disease trials and pediatric studies will be discussed, where the focus will be on methodological and practical considerations.

### ASIED: a Bayesian adaptive subgroup-identification enrichment design

Yanxun Xu<sup>1</sup>, Florica Constantine<sup>1</sup>, ♦Yuan Yuan<sup>2</sup> and Yili Pritchett<sup>3</sup>

<sup>1</sup>Johns Hopkins University

<sup>2</sup>AstraZeneca

<sup>3</sup>Biometrics, G1 Therapeutics, Inc  
yuan.yuan2@astrazeneca.com

Developing targeted therapies based on patients' baseline characteristics and genomic profiles such as biomarkers has gained growing interests in recent years. Depending on patients' clinical characteristics, the expression of specific biomarkers or their combinations, different patient subgroups could respond differently to the same treatment. An ideal design, especially at the proof of concept stage, should search for such subgroups and make dynamic adaptation as the trial goes on. When no prior knowledge is available on whether the treatment works on the all-comer population or only works on the subgroup defined by one biomarker or several biomarkers, it is necessary to incorporate the adaptive estimation of the heterogeneous treatment effect to the decision-making at in-

terim analyses. To address this problem, we propose an Adaptive Subgroup-Identification Enrichment Design, ASIED, to simultaneously search for predictive biomarkers, identify the subgroups with differential treatment effects, and modify study entry criteria at interim analyses when justified. More importantly, we construct robust quantitative decision-making rules for population enrichment when the interim outcomes are heterogeneous in the context of a multilevel target product profile, which defines the minimal and targeted levels of treatment effect. Through extensive simulations, the ASIED is demonstrated to achieve desirable operating characteristics and compare favorably against alternatives.

### **Innovative Adaptive Designs for Investigational Drug Development in Small Populations: A Discussion of Case Studies**

*Godwin Yung*

Genentech

yunggg@gene.com

In areas of unmet need, conducting fully powered randomized controlled trials (RCTs) may be infeasible due to enrollment challenges, timeline constraints, cost, or sometimes ethical concerns. As part of an effort to modernize approaches for drug development, in November 2019, FDA finalized the guidance for adaptive designs. This guidance emphasizes four principles for designing, conducting, and reporting results from adaptive clinical trials. It also provides advice on what information should be submitted for Bayesian and other complex trial designs. Moreover, FDA is currently conducting a 5-year Complex Innovative Trial Design (CID) Pilot Meeting Program to support the goal of facilitating and advancing the use of complex clinical trials through June 2022. With the advancement of technology, information available from sources of real-world data (RWD) have grown rapidly. Under the framework of adaptive clinical trial designs, such information can be utilized to generate evidence for making efficacy claim during interim or final analysis. Lab and genetic tests with quick turnaround can be employed to personalize treatment for patients in ongoing studies. Especially for drug development targeting small/vulnerable populations (e.g., rare diseases or children), unique challenges such as small patient numbers, heterogeneity in disease presentation, and a lack of understanding of the mechanism, call for more innovative thinking in the design and analysis of trials. For instance, how can the variability from historical data and the current trial data be accounted for during an interim analysis for different data types? How can external data from heterogeneous populations be incorporated? Are there different considerations when it comes to single-arm or multiple-arm trials? In this presentation, a few examples of rare disease trials and pediatric studies will be discussed, where the focus will be on methodological and practical considerations.

### **INSIGHT: A Bayesian Adaptive Platform Trial to Develop Precision Medicines for Patients With Glioblastoma**

*Lorenzo Trippa*

Dana-Farber Cancer Institute

ltrippa@jimmy.harvard.edu

Adequately prioritizing the numerous therapies and biomarkers available in late-stage testing for patients with glioblastoma (GBM) requires an efficient clinical testing platform. We developed and implemented INSIGHT (Individualized Screening Trial of Innovative Glioblastoma Therapy) as a novel adaptive platform trial (APT) to develop precision medicine approaches in GBM.

### **Session 79: Advanced Statistical Learning for High-dimensional Heterogeneous Data**

#### **Random projection pursuit regression**

*Qichen Liao<sup>1</sup>, Wei Zhang<sup>1</sup>, Jian Guo<sup>2</sup> and ♦Sijian Wang<sup>3</sup>*

<sup>1</sup>Tsinghua University

<sup>2</sup>International Digital Economy Academy

<sup>3</sup>Rutgers University

sijian.wang@rutgers.edu

Projection pursuit regression (PPR) adapts the additive models in that it first projects the data matrix of explanatory variables in the optimal direction before applying smoothing functions to these explanatory variables. As a consequence of this, PPR can be quite flexible to approximate a complicated regression function. In this talk, we draw some connections between PPR and boosting and neuron networks. Motivated by these connections and borrowing the spirit of random forest, we introduce a random projection pursuit regression. The proposed method can be more flexible than PPR in terms of capturing the relationship between outcome and covariates. It can also be more efficient (requiring smaller sample size) than deep neuron networks in terms of recognizing patterns and may have better prediction performance than deep neuron networks when the sample size is moderate. The method is demonstrated with both simulation studies and real data analysis.

#### **Subgroup Inference for Heterogeneous Treatment Effect Estimation**

*♦Lu Tang<sup>1</sup> and Ling Zhou<sup>2</sup>*

<sup>1</sup>University of Pittsburgh

<sup>2</sup>Southwestern University of Finance and Economics

lutang@pitt.edu

INSIGHT compares experimental arms with a common control of standard concurrent temozolomide and radiation therapy followed by adjuvant temozolomide. The primary end point is overall survival. Patients with newly diagnosed unmethylated GBM who are IDH R132H mutation negative and with genomic data available for biomarker grouping are eligible. At the initiation of INSIGHT, three experimental arms (neratinib, abemaciclib, and CC-115), each with a proposed genomic biomarker, are tested simultaneously. Initial randomization is equal across arms. As the trial progresses, randomization probabilities adapt on the basis of accumulating results using Bayesian estimation of the biomarker-specific probability of treatment impact on progression-free survival. Treatment arms may drop because of low probability of treatment impact on overall survival, and new arms may be added. Detailed information on the statistical model and randomization algorithm is provided to stimulate discussion on trial design choices more generally and provide an example for other investigators developing APTs.

#### **Joint Robust Multiple Inference on Large-Scale Multivariate Regression**

*♦Wen Zhou<sup>1</sup>, Youngseok Song<sup>1</sup> and Wenxin Zhou<sup>2</sup>*

<sup>1</sup>Colorado State University

<sup>2</sup>University of California, San Diego

rickzhouwen@gmail.com

Large scale multivariate regression with many heavy-tailed responses arises in a wide range of areas from genomics, financial asset pricing, banking regulation, to psychology and social studies. Simultaneously testing a large number of general linear hypotheses, such as multiple contrasts, based on the large scale multivariate regression reveals a variety of associations between responses and regression or experimental factors. Traditional multiple testing methods often ignore the effect of heavy-tailedness in the data and im-

pose joint normality assumption that is arguably stringent in applications. This results in unreliable conclusions due to the loss of control on the false discovery rate and severe compromise of power in practice. In this paper, we employ data-adaptive Huber regression to propose a framework of joint robust inference of the general linear hypotheses for large scale multivariate regression. With mild conditions, we show that the proposed method produces consistent estimate of the false discovery proportion and the false discovery rate at a pre-specified level. Particularly, we employ a bias-correction robust covariance estimator and study its exponential-type deviation inequality to provide theoretical guarantee of our proposed multiple testing framework. Extensive numerical experiments demonstrate the gain in power of the proposed method compared to the ordinary least square and other procedures.

### Multicategory Angle-based Learning for Estimating Optimal Dynamic Treatment Regimes with Censored Data

◆ *Fei Xue*<sup>1</sup>, *Yanqing Zhang*<sup>2</sup>, *Wenzhuo Zhou*<sup>3</sup>, *Haoda Fu*<sup>4</sup> and *Annie Qu*<sup>5</sup>

<sup>1</sup>University of Pennsylvania

<sup>2</sup>Yunnan University

<sup>3</sup>University of Illinois at Urbana-Champaign

<sup>4</sup>Eli Lilly and Company

<sup>5</sup>University of California Irvine  
xuefei012@gmail.com

An optimal dynamic treatment regime (DTR) consists of a sequence of decision rules in maximizing long-term benefits, which is applicable for chronic diseases such as HIV infection or cancer. In this paper, we develop a novel angle-based approach to search the optimal DTR under a multicategory treatment framework for survival data. The proposed method targets to maximize the conditional survival function of patients following a DTR. In contrast to most existing approaches which are designed to maximize the expected survival time under a binary treatment framework, the proposed method solves the multicategory treatment problem given multiple stages for censored data. Specifically, the proposed method obtains the optimal DTR via integrating estimations of decision rules at multiple stages into a single multicategory classification algorithm without imposing additional constraints, which is also more computationally efficient and robust. In theory, we establish Fisher consistency and provide the risk bound for the proposed estimator under regularity conditions. Our numerical studies show that the proposed method outperforms competing methods in terms of maximizing the conditional survival probability. We apply the proposed method to two real datasets: Framingham heart study data and acquired immunodeficiency syndrome (AIDS) clinical data.

### Session 80: New development in Bayesian methods and algorithms

#### Kriging: Beyond Matérn

*Anindya Bhadra*

Purdue University  
bhadra@purdue.edu

The Matérn covariance function is a popular choice for prediction in spatial statistics and uncertainty quantification literature. A key benefit of the Matérn class is that it is possible to get precise control over the degree of differentiability of the process realizations. However, the Matérn class possesses exponentially decaying tails, and thus may not be suitable for modeling polynomially decaying dependence. This problem can be remedied using polynomial covari-

ances; however one loses control over the degree of differentiability of the process realizations, in that the realizations using polynomial covariances are either infinitely differentiable or not differentiable at all. We construct a new family of covariance functions called the emphConfluent Hypergeometric (CH) class using a scale mixture representation of the Matérn class where one obtains the benefits of both Matérn and polynomial covariances. The resultant covariance contains two parameters: one controls the degree of differentiability near the origin and the other controls the tail heaviness, independently of each other. Using a spectral representation, we derive theoretical properties of this new covariance including equivalence measures and asymptotic behavior of the maximum likelihood estimators under infill asymptotics. The improved theoretical properties in predictive performance of this new covariance class are verified via extensive simulations. Application using NASA's Orbiting Carbon Observatory-2 satellite data confirms the advantage of this new covariance class over the Matérn class, especially in extrapolative settings.

### A Bayesian finite mixture model with variable selection for data with mixed-type variables

◆ *Shu Wang*<sup>1</sup> and *Chung-Chou Chang*<sup>2</sup>

<sup>1</sup>University of Florida

<sup>2</sup>University of Pittsburgh  
swang0221@ufl.edu

Finite mixture model is an important branch of clustering methods and can be used for data sets with mixed types of variables. We extend this modeling approach by incorporating the feature of variable selection and handling biomarkers subject to limits of detection (LOD) using Bayesian approach. Specifically, we use spike-and-slab prior for categorical variables so that variable selection for mixed data is feasible and use a Bayesian approach in estimation to bypass the limitation of the EM algorithm. To handle biomarkers subject to LOD, we iteratively fill in censored values in the Gibbs sampling iterations. Simulations across various scenarios were conducted to examine the performance of our proposed model. We applied our method to identify sepsis phenotypes among patients admitted to intensive care units from a health care system.

### Ultra-Fast Approximate Inference Using Variational Functional Mixed Models

*Shuning Huo*<sup>1</sup>, *Jeffrey S Morris*<sup>2</sup> and ◆ *Hongxiao Zhu*<sup>1</sup>

<sup>1</sup>Virginia Tech

<sup>2</sup>University of Pennsylvania  
hongxiao@vt.edu

While Bayesian functional mixed models have been shown effective to model functional data with various complex structures, their application to extremely high-dimensional data is limited due to computational challenges involved in posterior sampling. We introduce a new computational framework that enables ultra-fast approximate inference for high-dimensional data in functional form. This framework adopts parsimonious basis to represent functional observations, which facilitates efficient compression and parallel computing in basis space. Instead of performing expensive Markov chain Monte Carlo sampling, we approximate the posterior distribution using variational Bayes and adopt fast iterative algorithms to estimate parameters of the approximate distribution. Our approach provides a multiple testing procedure in basis space, which can be used to identify significant local regions on functional data that are nonzero or different across groups of samples. We perform two simulation studies to assess the accuracy of approximate inference and the computational benefits, and demonstrate applications of the pro-

posed approach by using two real datasets—the mass spectrometry data in a cancer study and the brain imaging data in a Alzheimer’s disease study.

### Recent Advances in Bayesian Methods for the Analysis of Tumor Pathology Images

◆ Qiwei Li<sup>1</sup>, Cong Zhang<sup>1</sup> and Guanghua Xiao<sup>2</sup>

<sup>1</sup>The University of Texas at Dallas

<sup>2</sup>The University of Texas Southwestern Medical Center  
qiwei.li@utdallas.edu

With the advance of imaging technology, digital pathology imaging of tumor tissue slides is becoming a routine clinical procedure for cancer diagnosis. This process produces massive imaging data that capture histological details in high resolution. Recent developments in deep-learning methods have enabled us to classify individual regions and cells from digital pathology images on a large scale. Reliable statistical approaches to model the resulting data at the histological and cellular levels are urgently needed, as they can provide new insight into tumor progression and shed light on the biological mechanisms of cancer. From the identified tumor regions, we extract the tumor boundary and consider it as a closed polygonal chain. A novel Bayesian model is proposed to identify landmark points of the chain to provide descriptive statistics and characterize tumor boundary roughness. From the identified commonly seen cells (i.e. lymphocyte, stromal, and tumor), we collect their spatial locations and consider them from a marked point process. Two novel spatial models with interpretable underlying parameters are proposed in a Bayesian framework to quantify their homogeneous and heterogeneous spatial correlations, respectively. Several case studies are conducted on the pathology images of 188 lung cancer patients from the National Lung Screening Trial to demonstrate those clinically informative features defined by the above Bayesian models. Our Bayesian models can also be used to characterize the shape and spatial features in many other fields.

### Session 81: Innovative methods for complex censored data

#### Infinite Parameter Estimates in Proportional Hazards Regression

John Kolassa<sup>1</sup> and ◆ Jane Zhang<sup>2</sup>

<sup>1</sup>Rutgers University

<sup>2</sup>Abbvie  
zhang\_jane@allergan.com

Proportional hazards are used to model event time data subject to censoring. Inference is performed using a partial likelihood, which has many properties in common with the full likelihood and is generally used in the same way. For example, parameters are often estimated by maximizing the partial likelihood, standard errors are often calculated from the second derivative of the log of the partial likelihood functions, and the change in the maximized likelihood as one moves from a larger model to a smaller model nested within it is often used for testing. Small samples involving discrete covariates with strong effects can lead to infinite maximum partial likelihood estimates. The monotonicity in the likelihood complicates statistical inference. A naive response to the problem of likelihood monotonicity is to report numerical results at the last computable step, and to exit with a warning message. A more sophisticated approach involves regularization or a penalty function. A methodology is presented for eliminating nuisance parameters estimated at infinity using approximate conditional inference.

#### Recent development on hierarchical joint frailty models for joint modeling of recurrent events and a terminal event

Zheng Li<sup>1</sup>, Vern M. Chinchilli<sup>2</sup> and ◆ Ming Wang<sup>2</sup>

<sup>1</sup>Novartis

<sup>2</sup>Penn State College of Medicine  
mwang@phs.psu.edu

Recurrent events could be stopped by a terminal event, which commonly occurs in biomedical and clinical studies. Taking the Cardiovascular Health Study (CHS) as a motivating example, patients can experience recurrent events of myocardial infarction (MI) or stroke during follow-up, which, however, can be truncated by death. Since death could be a devastating complication of myocardial infarction or stroke recurrences, ignoring dependent censoring when analyzing recurrent events may lead to invalid inference. The joint frailty model is widely used but with several limitations, such as the assumption of conditional independence, constant correlation between recurrent events and death and so on. Thus, hierarchical joint frailty models have been recently developed for more valid and informative inference. We will focus on potential model extensions under different scenarios and emphasis on the predicted accuracy assessment of this method. Extensive simulation studies are performed for evaluation to enhance its applications, and lastly illustrated by the CHS study.

#### Conditional association and concordance for bivariate censored data

Ruoshan Li

The University of Texas Health Science Center at Houston  
ruoshan.li@uth.tmc.edu

When analyzing bivariate outcome data, it is often of scientific interest to measure and estimate the association between the bivariate outcomes. In the presence of influential covariates for one or both of the outcomes, conditional association measures can quantify the strength of association without the disturbance of the marginal covariate effects, to provide cleaner and less-confounded insights into the bivariate association. In this work, we propose estimation and inferential procedures for assessing the conditional Kendall’s tau coefficient given the covariates, by adopting the quantile regression and quantile copula framework to handle marginal covariate effects. The proposed method can flexibly accommodate right censoring and be readily applied to bivariate survival data. It also facilitates an estimator of the conditional concordance measure, namely, a conditional C index, where the unconditional C index is commonly used to assess the predictive capacity for survival outcomes. The proposed method is flexible and robust and can be easily implemented using standard software. Application of our methods to a real-life data example demonstrates their desirable practical utility.

#### Regression with Covariates subject to Limits of Detection

Jimmy Kwon and ◆ Bin Nan

UC Irvine  
nanb@uci.edu

We consider generalized linear models with left-censored covariates due to the lower limits of detection. It has been shown that the complete case analysis by eliminating observations with values below limit of detection yields valid estimates for regression coefficients, but loses efficiency; substitution methods are biased; and maximum likelihood method relies on parametric models for the unobservable tail probability, thus may suffer from model misspecification. To obtain robust and more efficient results, a semiparametric pseudo-likelihood approach for the regression parameters using an accelerated failure time model for a single covariate subject to limit of detection was considered in the literature, and a two-stage estimation



procedure was proposed, where the conditional distribution of the covariate with limit of detection given other variables is estimated prior to maximizing the likelihood function for the regression parameters. When there are two or more covariates subject to their limits of detection, however, the implementation of the two-stage semiparametric method becomes much more difficult. The added challenge will be discussed in this talk.

## Session 82: Student Paper Award Invited Session

### BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data

♦ *Xinjun Wang<sup>1</sup>, Zhe Sun<sup>1</sup>, Yanfu Zhang<sup>1</sup>, Zhongli Xu<sup>1</sup>, Hongyi Xin<sup>1</sup>, Heng Huang<sup>1</sup>, Richard Duerr, Kong Chen<sup>1</sup>, Ying Ding<sup>1</sup> and Wei Chen<sup>1</sup>*

<sup>1</sup>University of Pittsburgh  
xiw119@pitt.edu

Droplet-based single cell transcriptome sequencing (scRNA-seq) technology, largely represented by the 10× Genomics Chromium system, is able to measure the gene expression from tens of thousands of single cells simultaneously. More recently, coupled with the cutting-edge Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq), the droplet-based system has allowed for immunophenotyping of single cells based on cell surface expression of specific proteins together with simultaneous transcriptome profiling in the same cell. Despite the rapid advances in technologies, novel statistical methods and computational tools for analyzing multi-modal CITE-Seq data are lacking. In this study, we developed BREM-SC, a novel Bayesian Random Effects Mixture model that jointly clusters paired single cell transcriptomic and proteomic data. Through simulation studies and analysis of public and in-house real data sets, we successfully demonstrated the validity and advantages of this method in fully utilizing both types of data to accurately identify cell clusters. In addition, as a probabilistic model-based approach, BREM-SC is able to quantify the clustering uncertainty for each single cell. This new method will greatly facilitate researchers to jointly study transcriptome and surface proteins at the single cell level to make new biological discoveries, particularly in the area of immuno

### Bayesian Meta-analysis of Censored Rare Events with Stochastic Coarsening

♦ *Xinyue Qi<sup>1</sup>, Christine B. Peterson<sup>2</sup>, Yucai Wang<sup>3</sup> and Shouhao Zhou<sup>4</sup>*

<sup>1</sup>The University of Texas Health Science Center at Houston

<sup>2</sup>The University of Texas MD Anderson Cancer Center

<sup>3</sup>Mayo Clinic

<sup>4</sup>Pennsylvania State University  
xinyue.qi.2020@gmail.com

Meta-analysis is a powerful tool for drug safety assessment by synthesizing findings from independent clinical trials. However, published clinical studies may or may not report all adverse events (AEs) if the observed number of AEs were fewer than a pre-specified study-dependent cutoff, which can be considered as a special type of data coarsening (Heitjan and Rubin, 1991) for rare events. To derive exact and robust inference, we investigate the stochastic nature of informative censoring and the conditions of ignorability for coarsened data mechanism. The proposed approach is illustrated using data from a recent meta-analysis of 125 clinical trials involving PD-1/PD-L1 inhibitors with respect to their toxicity profiles. We demonstrate that if the censored information is ignored, the incidence probability of AE is overestimated; this bias

could have significant impact on immunotherapy drug adoption and public health policy.

### Inference for BART with Multinomial Outcomes

♦ *Yizhen Xu<sup>1</sup> and Joseph Hogan<sup>2</sup>*

<sup>1</sup>Johns Hopkins University

<sup>2</sup>Brown University

yizhen.xu@alumni.brown.edu

The multinomial probit Bayesian additive regression trees (MP-BART) framework was proposed by Kindo et al. (2016) to approximate the latent utilities in multinomial probit (MNP) model with Bayesian additive regression trees (BART). Compared to multinomial logistic models, MNP does not assume independent alternatives and naturally obtain a Bayesian estimation of the correlation structure among alternatives through the multivariate Gaussian distributed latent utilities. We introduce two algorithms for fitting the MPBART and show that the theoretical mixing rates of our proposals are at least as good as the existing algorithm (KD) proposed by Kindo et al. We also discuss the robustness of the methods to the choice of reference level, the imbalance in outcome frequencies, and the specifications of prior hyperparameters for the utility error term. Through simulations and application, we observe improvement in our proposals compared to KD in terms of MCMC convergence rate and posterior predictive accuracy.

### Covariate Adaptive Family-wise Error Rate Control for Genome-Wide Association Studies

♦ *Huijuan Zhou<sup>1</sup>, Xianyang Zhang<sup>2</sup> and Jun Chen<sup>3</sup>*

<sup>1</sup>Renmin University of China and Texas A&M University

<sup>2</sup>Texas A&M University

<sup>3</sup>Mayo Clinic

huijuan@stat.tamu.edu

The family-wise error rate (FWER) has been widely used in genome-wide association studies. With the increasing availability of functional genomics data, it is possible to increase the detection power by leveraging these genomic functional annotations. Previous efforts to accommodate covariates in multiple testing focus on the false discovery rate control while covariate-adaptive FWER-controlling procedures remain under-developed. Here we propose a novel covariate-adaptive FWER-controlling procedure that incorporates external covariates which are potentially informative of either the statistical power or the prior null probability. An efficient algorithm is developed to implement the proposed method. We prove its asymptotic validity and obtain the rate of convergence through a perturbation-type argument. Our numerical studies show that the new procedure is more powerful than competing methods and maintains robustness across different settings. We apply the proposed approach to the UK Biobank data and analyze 27 traits with 9 million single-nucleotide polymorphisms tested for associations. Seventy-five genomic annotations are used as covariates. Our approach detects more genome-wide significant loci than other methods in 21 out of the 27 traits

### Functional Group Bridge for Simultaneous Regression and Support Estimation

♦ *Zhengjia Wang<sup>1</sup>, John Magnotti<sup>2</sup>, Michael Beauchamp<sup>2</sup> and Meng Li<sup>1</sup>*

<sup>1</sup>Rice University

<sup>2</sup>Baylor College of Medicine

zw23@rice.edu

There is a wide variety of applications where the unknown nonparametric functions are locally sparse, and the support of a function as well as the function itself is of primary interest. In the function-

on-scalar setting, while there has been a rich literature on function estimation, the study of recovering the function support is limited. In this article, we consider the function-on-scalar mixed effect model and propose a weighted group bridge approach for simultaneous function estimation and support recovery, while accounting for heterogeneity present in functional data. We use B-splines to transform sparsity of functions to its sparse vector counterpart of increasing dimension, and propose a fast non-convex optimization algorithm for estimation. We show that the estimated coefficient functions are rate optimal in the minimax sense under the L2 norm and resemble a phase transition phenomenon. For support estimation, we derive a convergence rate under the  $L_\infty$  norm that leads to a sparsistency property under  $\delta$ -sparsity, and provide a simple sufficient regularity condition under which a strict sparsistency property is established. An adjusted extended Bayesian information criterion is proposed for parameter tuning. The developed method is illustrated through simulation and an application to a novel intracranial electroencephalography (iEEG) dataset.

### Session 83: Statistical Methods for design and analysis of health outcomes

#### The effect of biomarker variability on clinical outcomes: comparing different methods

♦ *Feng Gao, Jingqin Luo, Jinxia Liu, Chengjie Xiong and Lei Liu*  
Washington University School of Medicine  
feng@wustl.edu

As the field of precision medicine continues to advance, methodologies that utilize individual patient information over time are becoming increasingly valuable to provide more accurate prognosis as early as possible. In many clinical trials and epidemiologic studies, information is gathered on both time to event (survival data) and repeated measures of biomarkers (longitudinal data). In some clinical trials and epidemiologic studies, investigators are interested in knowing whether the variability of a biomarker is independently predictive of clinical outcomes. This question is often addressed via a two-stage approach where a sample-based estimate (e.g., standard deviation) is calculated as a surrogate for the “true” variability and then used in regression models as a covariate assumed to be free of measurement error. However, it is well known that the measurement error in covariates could cause underestimation of the true association in such a two-stage approach. The issue of underestimation can be substantial when the precision is low because of limited number of measures per subject. This talk will compare several different two-stage methods in assessing the effect of biomarker variability on time-to-event outcome. The methods will be compared using data simulated from a joint model of longitudinal and survival data. The methods will also be illustrated in the Ocular Hypertension Treatment Study to assess whether the variability of intraocular pressure is an independent risk of primary open-angle glaucoma.

#### Optimal designs in three-level cluster randomized trials with a binary outcome

♦ *Jingxia Liu<sup>1</sup>, Lei Liu<sup>2</sup> and Graham Colditz<sup>1</sup>*

<sup>1</sup>Washington University School of Medicine

<sup>2</sup>Washington University School of Medicine (WUSM)  
esther@wustl.edu

Cluster randomized trials (CRTs) were originally proposed for use when randomization at the subject level is practically infeasible or may lead to a severe estimation bias of the treatment effect. However, recruiting an additional cluster costs more than enrolling an

additional subject in an individually randomized trial. Under budget constraints, researchers have proposed the optimal sample sizes in two-level CRTs. CRTs may have a three-level structure, in which two levels of clustering should be considered. In this paper, we propose optimal designs in three-level CRTs with a binary outcome, assuming nested exchangeable correlation structure in generalized estimating equation models. We provide the variance of estimators of three commonly used measures: risk difference, risk ratio, and odds ratio. For a given sampling budget, we discuss how many clusters and how many subjects per cluster are necessary to minimize the variance of each measure estimator. For known association parameters, the locally optimal design (LOD) is proposed. When association parameters are unknown but within pre-determined ranges, the MaxiMin design (MMD) is proposed to maximize the minimum of relative efficiency over the possible ranges, that is, to minimize the risk of the worst scenario.

#### Matched or unmatched analyses with propensity-score-matched data?

*Fei Wan*

Washington university in St. Louis  
wan.fei@wustl.edu

Propensity-score matching has been used widely in observational studies to balance confounders across treatment groups. However, whether matched-pairs analyses should be used as a primary approach is still in debate. We compared the statistical power and type 1 error rate for four commonly used methods of analyzing propensity-score-matched samples with continuous outcomes: (1) an unadjusted mixed-effects model, (2) an unadjusted generalized estimating method, (3) simple linear regression, and (4) multiple linear regression. Multiple linear regression had the highest statistical power among the four competing methods. We also found that the degree of intraclass correlation within matched pairs depends on the dissimilarity between the coefficient vectors of confounders in the outcome and treatment models. Multiple linear regression is superior to the unadjusted matched-pairs analyses for propensity-score-matched data.

#### Regression analysis of clustered failure time data with informative cluster size under the additive transformation models

♦ *Ling Chen<sup>1</sup>, Yanqin Feng<sup>2</sup> and Jianguo Sun<sup>3</sup>*

<sup>1</sup>Washington University in St Louis School of Medicine

<sup>2</sup>Wuhan University

<sup>3</sup>University of Missouri  
lingchen@wustl.edu

This paper discusses regression analysis of clustered failure time data, which occur when the failure times of interest are collected from clusters. In particular, we consider the situation where the correlated failure times of interest may be related to cluster sizes. For inference, we present two estimation procedures, the weighted estimating equation-based method and the within-cluster resampling-based method, when the correlated failure times of interest arise from a class of additive transformation models. The former makes use of the inverse of cluster sizes as weights in the estimating equations, while the latter can be easily implemented by using the existing software packages for right-censored failure time data. An extensive simulation study is conducted and indicates that the proposed approaches work well in both the situations with and without informative cluster size. They are applied to a dental study that motivated this study.

## Session 84: Statistical Modeling for COVID-19

### Curating A COVID-19 Data Repository and Forecasting County-Level Death Counts in the United States

Nick Altieri<sup>1</sup>, Rebecca L. Barter<sup>1</sup>, James Duncan<sup>1</sup>, Raaz Dwivedi<sup>1</sup>, Karl Kumbier<sup>2</sup>, Xiao Li<sup>1</sup>, Robert Netzorg<sup>1</sup>, Briton Park<sup>1</sup>, Chandan Singh<sup>1</sup>, Yan Shuo Tan<sup>1</sup>, Tiffany Tang<sup>1</sup>, Yu Wang<sup>1</sup>, Chao Zhang<sup>1</sup> and ♦Bin Yu<sup>1</sup>

<sup>1</sup>University of California, Berkeley

<sup>2</sup>University of California, San Francisco  
binyu@berkeley.edu

As the COVID-19 outbreak evolves, accurate forecasting continues to play an extremely important role in informing policy decisions. In this paper, we present our continuous curation of a large data repository containing COVID-19 information from a range of sources. We use this data to develop predictions and corresponding prediction intervals for the short-term trajectory of COVID-19 cumulative death counts at the county-level in the United States up to two weeks ahead. Using data from January 22 to June 20, 2020, we develop and combine multiple forecasts using ensembling techniques, resulting in an ensemble we refer to as Combined Linear and Exponential Predictors (CLEP). Our individual predictors include county-specific exponential and linear predictors, a shared exponential predictor that pools data together across counties, an expanded shared exponential predictor that uses data from neighboring counties, and a demographics-based shared exponential predictor. We use prediction errors from the past five days to assess the uncertainty of our death predictions, resulting in generally-applicable prediction intervals, Maximum (absolute) Error Prediction Intervals (MEPI). MEPI achieves a coverage rate of more than 94% when averaged across counties for predicting cumulative recorded death counts two weeks in the future. Our forecasts are currently being used by the nonprofit organization, Response4Life, to determine the medical supply need for individual hospitals and have directly contributed to the distribution of medical supplies across the country. We hope that our forecasts and data repository (available upon request) can help guide necessary county-specific decision-making and help counties prepare for their continued fight against COVID-19.

### Semiparametric Bayesian Inference for the Transmission Dynamics of COVID-19 with a State-Space Model

Tianjian Zhou<sup>1</sup> and ♦Yuan Ji<sup>2</sup>

<sup>1</sup>Colorado State University

<sup>2</sup>The University of Chicago  
yji@health.bsd.uchicago.edu

The outbreak of Coronavirus Disease 2019 (COVID-19) is an ongoing pandemic affecting over 200 countries and regions. Inference about the transmission dynamics of COVID-19 can provide important insights into the speed of disease spread and the effects of mitigation policies. We develop a novel Bayesian approach to such inference based on a probabilistic compartmental model using data of daily confirmed COVID-19 cases. In particular, we consider a probabilistic extension of the classical susceptible-infectious-recovered model, which takes into account undocumented infections and allows the epidemiological parameters to vary over time. We estimate the disease transmission rate via a Gaussian process prior, which captures nonlinear changes over time without the need of specific parametric assumptions. We utilize a parallel-tempering Markov chain Monte Carlo algorithm to efficiently sample from the highly correlated posterior space. Predictions for future observations are done by sampling from their posterior predictive distributions.

Performance of the proposed approach is assessed using simulated datasets. Finally, our approach is applied to COVID-19 data from four states of the United States: Washington, New York, California, and Illinois. An R package BaySIR is made available at this [https](https://github.com/raazdwivedi/baySIR) URL for the public to conduct independent analysis or reproduce the results in this paper.

### Transmission dynamic modeling for COVID-19 data in U.S. and the world

♦Haoyu Zhang<sup>1</sup>, Chaolong Wang<sup>2</sup> and Xihong Lin<sup>1</sup>

<sup>1</sup>Harvard T.H. Chan School of Public Health

<sup>2</sup>Huazhong University of Science and Technology  
haoyuzhang@hsph.harvard.edu

Understanding COVID-19 transmission dynamics, such as the temporal trend of the transmission rate, i.e., time-varying effective reproductive numbers ( $R_t$ ), is important for controlling the disease, estimating the prevalence and the total number of infections, and the fatality. Modeling COVID-19 transmission dynamics faces significant challenges. A large number of cases are likely to be unascertained/undetected due to insufficient testing. Many of these un-ascertained cases are asymptomatic and mildly symptomatic but still infectious. For a given region, the unknown ascertainment rate often varies over time due to evolving testing capacity. Furthermore, case counts are often subject to delayed reporting. Epidemic models that ignore unascertained cases and reporting delays can lead to biased estimates of the transmission rate and the prevalence of the disease. To address these problems, we develop an over-dispersed Poisson Partial Differential Equation transmission dynamic model to estimate the region-specific  $R_t$  and ascertainment rates as piecewise constants by (1) using daily case counts of positive tests and accounting for both reporting delay and over-dispersion; (2) incorporating daily death count data; (3) modeling the compartments of isolation at home and at hospitals separately. We also estimate the prevalence and the daily numbers of total infections in a given region by accounting for under-ascertainment. We apply the model to 50 U.S. states and more than 200 countries using data up to August 31st, 2020. Our results show that ascertainment rates continue to increase as the testing capacity increase. However, there is still a large proportion of unascertained COVID-19 cases.

### A spatiotemporal model for county-level covid-19 infection data in the USA

Peter Song

University of Michigan  
pxsong@umich.edu

We develop a health information system that provides micro COVID-19 infection risk predictions in a similar way to the weather forecast. Generalizing the von Neumann's cellular automata with the stochastic infectious disease models, this forecast system integrates multiple sources of information in the risk prediction, including serological test surveys, mobile device data and personal mobility scores. This prediction model is tuned by minimizing the prediction error of one-day ahead projection of infection prevalence. We illustrate the proposed method by the county-level COVID-19 prevalence projection over 3,109 counties in the continental US. In comparison to the conventional temporal risk prediction models, our model informs high-resolution community-level risk projection, which is useful for tailored decision-making on business reopenings and medical resource allocation.

## Session 85: Recent statistical advancements in the design of clinical trials

### Innovative Group sequential Design with Optimal Futility Stopping rules

♦ *Zhaoyang Teng<sup>1</sup>, Qiang Zhao<sup>2</sup>, Rui Tang<sup>1</sup> and Yi Liu<sup>2</sup>*

<sup>1</sup>Servier Pharmaceuticals

<sup>2</sup>Nektar Therapeutics

[zhaoyang.teng@servier.com](mailto:zhaoyang.teng@servier.com)

Group sequential design (GSD) has been widely used in pharmaceutical industry for drug development, especially in phase III pivotal studies. The efficacy boundaries usually can be determined by alpha spending function, such as O'Brien-Fleming, Pocock, etc. To determine the futility boundaries, the classical approach is to use  $\beta$ -spending function under efficacious treatment effect. However, there are two critical issues with this traditional approach. First, it does not give the same amount of consideration on the probability of stopping under an inefficacious treatment which is more clinically meaningful. Second, the final analysis is considered as the last futility checking while in fact it shouldn't be because the study either succeed or fail at that time. The ideal futility analysis in a GSD should be set up in a way that a trial can be stopped for futility with certain assurance before final analysis given an inefficacious treatment effect. We develop a framework to design futility analysis for a GSD, in which the futility boundaries are determined under inefficacious treatment effect using futility stopping probability (FSP) spending according to a prespecified interim analysis as futility final analysis instead of the trial's final analysis. With the proposed new framework, the futility boundaries are more interpretable with a clear goal set up at design stage and smaller sample size is needed to achieve the same study power using FSP compared with traditional GSD with  $\beta$ -spending function.

### A Bayesian Adaptive Design for Concurrent Trials Involving Biologically-Related Diseases

♦ *Tony Jiang<sup>1</sup>, Amy Xia<sup>1</sup>, Matthew Psioda<sup>2</sup>, Joseph Ibrahim<sup>2</sup> and Jiawei Xu<sup>2</sup>*

<sup>1</sup>Amgen

<sup>2</sup>UNC

[xunj@amgen.com](mailto:xunj@amgen.com)

We develop a Bayesian adaptive design framework for a clinical program where an investigational product is to be studied concurrently in a set of phase II trials for a set of biologically related diseases with the goal of demonstrating superiority to a control in each disease. The proposed approach borrows information on treatment effectiveness using Bayesian model averaging (BMA) to combine inference from analyses based on combinations of informative conjugate power priors that either strongly favor the null hypothesis or the hypothesized alternative in each disease indication. The approach allows for information borrowing without making untenable assumptions regarding how treatment effect parameters relate across the different diseases being studied and provides a powerful framework for information borrowing in a general setting where each disease has a potentially different endpoint (e.g., some binary and some continuous). Information borrowing is induced through elicited prior model probabilities that satisfy a dependence criterion which we show to be sufficient and necessary to result in information borrowing in the BMA context. We show via simulation that operating characteristics for trials based on the proposed design framework are favorable to those based on information borrowing designs using the Bayesian hierarchical model which is not well suited to the different endpoint problem.

### A method for sample size calculation via E-value in the planning of observational studies

♦ *Yixin Fang, Weili He, Xiaofei Hu and Hongwei Wang*

AbbVie

[yixin.fang@abbvie.com](mailto:yixin.fang@abbvie.com)

Confounding adjustment plays a key role in designing observational studies such as cross-sectional studies, case-control studies, and cohort studies. In this presentation, we propose a simple method for sample size calculation in observational research in the presence of confounding. The method is motivated by the notion of E-value, using some bounding factor to quantify the impact of confounders on the effect size. The method can be applied to calculate the needed sample size in observational research when the outcome variable is binary, continuous, or time-to-event. The method can be implemented straightforwardly using existing commercial software such as the PASS software. We demonstrate the performance of the proposed method through numerical examples, simulation studies, and a real application.

### Rethinking Treatment Switch in a randomized clinical trial (Innovative Design)

*Eiji Ishida*

FDA/CDER

[eiji.ishida@fda.hhs.gov](mailto:eiji.ishida@fda.hhs.gov)

This presentation illustrates the basic statistical elements in a scheme of treatment switch (crossover, stepped wedge, n of 1, sequential parallel comparison) used in a randomized study. As is well known, when treatment switch clinically well fits the purpose of evaluating efficacy or safety, positive correlations within each subject on an efficacy or safety endpoint of interest may lead to a great gain in efficiency, compared to a typical clinical trial with a parallel-group design. A crossover design or a design with treatment switch, however, may induce a non-identifiability problem caused by its over-parameterization. In a Williams design and a 2x2 crossover design, for instance, a treatment effect (defined as an efficacy difference between Drug and Placebo) may be infeasible to estimate in the presence of carryover effects. We look at a few examples of such designs used in a new drug development, and examine their model specifications and illustrate statistical issues through analyses of simulated data for the designs. We also illustrate a couple of designs in which an investigator may be able to escape a non-identifiability problem.

## Session 86: Challenging Statistical Issues in Oncology Studies

### Statistical Considerations on Using Minimal Residual Disease Status as the Efficacy Endpoint for Developing Novel Agents in Multiple Myeloma

♦ *Hong Tian, Jiajun Xu and Liang Xiu*

Johnson and Johnson

[htian@its.jnj.com](mailto:htian@its.jnj.com)

Treatment paradigm has shifted for the past 15 years in multiple myeloma. More and more effective agents become available, as a result, the progression free survival (PFS) rate has improved dramatically. It becomes impractical to evaluate novel agents using traditional endpoints within a reasonable timeframe. Minimal residual disease, which measures residual tumor cells at a deep level, is being identified as a putative surrogate endpoint for regulatory approvals. However, it remains important to size the confirmatory study according to the PFS endpoint to comply with regulatory requirements. Prediction of the treatment benefit on PFS endpoint

based on the observed effect on MRD is an interesting and important question to explore in the process of drug development. In this investigation, we explored a few practical options to empirically guestimate the relationship between the treatment effects on these two endpoints and evaluate the potential performance of these methods in sizing Phase 3 study using PFS endpoint based on observed MRD results in Phase 2 study.

### **An adaptive seamless phase II/III design with simultaneous treatment and subpopulation selections in clinical trials with survival endpoints**

*Cindy Lu<sup>1</sup> and ♦Liwen Wu<sup>2</sup>*

<sup>1</sup>Biogen

<sup>2</sup>University of Pittsburgh  
liw88@pitt.edu

The past decades have witnessed massive revolution of biomedical technology. Consequently these modern science has brought drug development to a new era, with pressing urge of innovation and creation from clinical trials, aiming to achieve shorter development cycle, lower costs and more importantly, improve patient care. With these regards, adaptive designs as an innovative design add more flexibility to the drug development paradigm. Seamless phase II/III designs, in particular, may help us reduce the development time of a drug. Two common application of seamless designs are adaptive treatment selection and adaptive subgroup enhancement designs. Our proposed design combines these two by performing simultaneous treatment and subgroup selection at the interim analysis for survival endpoint. We show that with carefully gauged decision cut-offs, our design is able to select the correct treatment and subgroup with high probabilities while keep the overall type I error rate controlled. The operating characteristics of the design are described by simulation studies.

### **Dynamic RMST Curves for Survival Analysis in Clinical Trials**

*♦Jason Liao, Frank Liu and Wen-Chi Wu*

Merck

jliao@incyte.com

The data from immuno-oncology therapy trials often show delayed effects, cure rate, crossing hazards, or some mixture of these phenomena. Thus, the common assumption of proportional hazards is often violated such that the commonly used log-rank test can be very underpowered. In these trials, the conventional hazard ratio for describing the treatment effect may not be a good estimand due to the lack of an easily understandable interpretation. To overcome this challenging issue, restricted mean survival time (RMST) has been strongly suggested and recommended for survival analysis in clinical literature due to its independence of the proportional hazard assumption and a more clinically meaningful interpretation. The RMST also aligns well with the estimand associated with the analysis from the recommendation in ICH E-9 (R1), and the test/estimation coherency. Currently, the Kaplan Meier (KM) curve is commonly applied to RMST related analyses. Due to some drawbacks of the KM approach such as the limitation in extrapolating to time points beyond follow-up time, and the large variance at time points with small numbers of events, the RMST may be hindered. To fully enhance the RMST method, the survival curve using a mixture model is proposed in this talk to construct dynamic RMST curves to evaluate and monitor survival analysis in clinical trials. This new RMST curve does not have the drawbacks of the KM approach. The good performance of this new proposal is illustrated through three real datasets. The proposed dynamic RMST curve can also be useful for checking whether follow-up time for

a study is long enough to demonstrate a treatment difference. The prediction feature of the dynamic RMST analysis may also be used for determining an appropriate time point for an interim analysis.

### **Analysis of Time to Event Data using a Flexible Mixture Model under a Constraint of Proportional Hazards**

*♦Frank Liu<sup>1</sup> and Jason Liao<sup>2</sup>*

<sup>1</sup>Merck & Co., Inc.

<sup>2</sup>Merck & Co., Inc.

guanghan.frank.liu@merck.com

Cox proportional hazards (PH) model evaluates effects of interested covariates under PH assumption without specified the baseline hazard. In clinical trial applications, however, the explicitly estimated hazard or cumulative survival function for each treatment group helps to assess and interpret the meaning of treatment difference. In this paper, we propose to use a flexible mixture model under the PH constraint to fit the underline survival functions. Simulations are conducted to evaluate its performance and show that the proposed mixture PH model is very similar to the Cox PH model in terms of estimating the hazard ratio, bias, confidence interval coverage, type-I error, and testing power. Application to several real clinical trial examples demonstrates that the results from this approach are almost identical to the results from Cox PH model. The explicitly estimated hazard function for each treatment group provides additional useful information and helps the interpretation of hazard comparisons.

## **Session 87: Recent advances in machine learning and causal inference**

### **Sparsity Double Robust Inference of Average Treatment Effects**

*Jelena Bradic<sup>1</sup>, Stefan Wager<sup>2</sup> and ♦Yinchu Zhu<sup>3</sup>*

<sup>1</sup>UCSD

<sup>2</sup>Stanford University

<sup>3</sup>University of Oregon

yzhu6@uoregon.edu

Many popular methods for building confidence intervals on causal effects under high-dimensional confounding require strong "ultra-sparsity" assumptions that may be difficult to validate in practice. To alleviate this difficulty, we here study a new method for average treatment effect estimation that yields asymptotically exact confidence intervals assuming that either the conditional response surface or the conditional probability of treatment allows for an ultra-sparse representation (but not necessarily both). This guarantee allows us to provide valid inference for average treatment effect in high dimensions under considerably more generality than available baselines. In addition, we showcase that our results are semi-parametrically efficient.

### **Graph Quilting**

*♦Giuseppe Vinci<sup>1</sup>, Gautam Dasarathy<sup>2</sup> and Genevera Allen<sup>3</sup>*

<sup>1</sup>University of Notre Dame

<sup>2</sup>Arizona State University

<sup>3</sup>Rice University

gvinci@nd.edu

Graphical model estimation and selection is a seemingly impossible task when several pairs of variables are never jointly observed. This unexplored statistical problem arises in several situations, such as in neuroimaging where, because of technology limitations, it is impossible to jointly record the activities of all neurons simultaneously. We call this statistical challenge the "Graph Quilting problem". In

the Gaussian graphical model, the unavailability of parts of the covariance matrix translates into the nonidentifiability of the precision matrix, which specifies the graph. However, we demonstrate that, under mild conditions, it is possible to correctly identify not only the edges connecting the observed pairs of nodes, but also a minimal superset of those connecting the variables that are never observed jointly. To accomplish the latter task, we devise a novel technique that we call the “Recursive-Complement” algorithm. We propose a graph estimator based on partially observed sample covariances and  $l_1$ -regularization, and establish its rates of convergence in high-dimensions. We illustrate the methodology using synthetic data, as well as data obtained from in vivo calcium imaging of ten thousand neurons in mouse visual cortex.

#### **Estimating the Effects of a New Technology using a Duration Model for Staggered Adoption (with Aureo de Paula (UCL))**

*Sida Peng*

Microsoft  
sidpeng@microsoft.com

Measuring the impact of new technologies on productivity has been a classic question in economics. However very little works have been focused on the revolution of the modern office place for knowledge workers. In this paper, we study the diffusion of Microsoft SharePoint, a new technology aimed to bring seamless collaboration for Microsoft Office documents, serving as a direct substitute for traditional email attachments. We propose a new type of staggered difference-in-difference estimator to measure the effect of SharePoint adoption. Through a parametric model of the adoption diffusion process, the proposed estimator allows us to account for latent variables that may modulate inter-temporal shocks before adoption. We show that our estimator works well in simulations and our empirical results show that the traditional email attachments can be reduced by 25 pieces per user after 10 months of adoption of SharePoint.

#### **Nonregular and Minimax Estimation of Individualized Thresholds in High dimension with Binary Responses**

*Yang Ning*

Cornell University  
yn265@cornell.edu

Given a large number of covariates  $Z$ , we consider the estimation of a high-dimensional parameter  $\theta$  in an individualized linear threshold  $\theta TZ$  for a continuous variable  $X$ , which minimizes the disagreement between  $\text{sign}(X - \theta TZ)$  and a binary response  $Y$ . While the problem can be formulated into the M-estimation framework, minimizing the corresponding empirical risk function is computationally intractable due to discontinuity of the sign function. Moreover, estimating  $\theta$  even in the fixed-dimensional setting is known as a nonregular problem leading to nonstandard asymptotic theory. To tackle the computational and theoretical challenges in the estimation of the high-dimensional parameter  $\theta$ , we propose an empirical risk minimization approach based on a regularized smoothed loss function. The statistical and computational trade-off of the algorithm is investigated. Statistically, we show that the finite sample error bound for estimating  $\theta$  in  $L_2$  norm is  $(\text{slog}d/n)\beta/(2\beta+1)$ , where  $d$  is the dimension of  $\theta$ ,  $s$  is the sparsity level,  $n$  is the sample size and  $\beta$  is the smoothness of the conditional density of  $X$  given the response  $Y$  and the covariates  $Z$ . The convergence rate is nonstandard and slower than that in the classical Lasso problems. Furthermore, we prove that the resulting estimator is minimax rate optimal up to a logarithmic factor. The Lepski's method is developed to achieve the adaption to the unknown sparsity  $s$  and smoothness  $\beta$ . Computa-

tionally, an efficient path-following algorithm is proposed to compute the solution path. We show that this algorithm achieves geometric rate of convergence for computing the whole path. Finally, we evaluate the finite sample performance of the proposed estimator in simulation studies and a real data analysis.

#### **Session 88: Recent advances in accounting for heterogeneity in complex data**

##### **A Statistical Framework for Genome-Scale Mutual Exclusivity Analysis of Cancer Mutations**

*Chi Wang*

University of Kentucky  
chi.wang@uky.edu

Cancer somatic driver mutations associated with genes within a pathway often show a mutually exclusive pattern across a cohort of patients. This mutually exclusive mutational signal has been frequently used to distinguish driver from passenger mutations and to investigate relationships among driver mutations. Current methods for de novo discovery of mutually exclusive mutational patterns are limited because the heterogeneity in background mutation rate can confound mutational patterns, and the presence of highly mutated genes can lead to spurious patterns. In addition, most methods only focus on a limited number of pre-selected genes and are unable to perform genome-wide analysis due to computational inefficiency. In this talk, we introduce a statistical framework, MEScan, for accurate and efficient mutual exclusivity analysis at the genomic scale. Our framework contains a fast and powerful statistical test for mutual exclusivity with adjustment of the background mutation rate and impact of highly mutated genes, and a multi-step procedure for genome-wide screening with the control of false discovery rate. We demonstrate that MEScan more accurately identifies mutually exclusive gene sets than existing methods and is at least two orders of magnitude faster than most methods. By applying MEScan to data from four different cancer types and pan-cancer, we have identified several biologically meaningful mutually exclusive gene sets.

##### **Dissecting high-throughput (epi)genomics signals from heterogeneous samples**

*Hao Wu*

Emory University  
hao.wu@emory.edu

Tissue samples obtained from clinical practices are usually mixtures of different cell types. The high-throughput data obtained from these samples are thus mixed signals. The cell mixture brings complications to data analysis, and will lead to biased results if not properly accounted for. In this talk, I will present some of our recent works on methods and strategies for analyzing high-throughput data from heterogeneous samples, including cell type-specific differential analysis and reference-free signal deconvolution.

##### **Efficient Gene-Environment Interaction Tests for Large Biobank-Scale Sequencing Studies with Correlated Samples**

♦ *Han Chen and Xinyu Wang*

The University of Texas Health Science Center at Houston  
han.chen.2@uth.tmc.edu

Complex human diseases are affected by both genetic and environmental risk factors. With the recent advance of sequencing technology, whole exome and genome sequencing data are now being collected at an unprecedented scale. However, large-scale sequencing studies (e.g., those involving biobanks) often include correlated samples, and existing statistical models accounting for relat-

edness, such as generalized linear mixed models (GLMMs), become computationally intensive in large samples. Here we propose efficient Mixed-model Association tests for GENE-Environment interactions (MAGEE), for testing gene-environment interactions (GEI) between an aggregate genetic variant set and environmental exposures on quantitative and binary traits in large biobank-scale sequencing studies. Joint tests for genetic main effects and GEI are also developed. A null GLMM adjusting for covariates but without any genetic effects is fitted only once in a whole genome GEI analysis, thereby vastly reducing the overall computational burden. Score tests for aggregate genetic variant sets are performed as a combination of genetic burden and variance component tests. The computational complexity is dramatically reduced in a whole genome GEI analysis, which makes MAGEE scalable to hundreds of thousands of individuals. Furthermore, MAGEE allows flexible weighting schemes to incorporate functional genomic information. We applied MAGEE to the exome sequencing data of related individuals from the UK Biobank as an illustrating example.

### Modeling the dynamics of disease transmission

♦ *Jing Huang and Jiasheng Shi*

University of Pennsylvania  
jinghuang0608@gmail.com

Reproduction number ( $R$ ), defined as the average number of people that will be infected by an individual who has the infection, plays a central role in predicting the evolution of an infectious disease outbreak. However, the  $R$  most certainly varies by location and by time due to multiple factors, including time-varying social distancing, varying public policies, difference in population density, and even changing of weather. To study the dynamics of disease transmission, we modeled the instantaneous reproduction number  $R_t$ ,  $t > 0$ , which can vary over time. Under the framework of quasi-score method, we proposed an online algorithm to iteratively estimate  $R_t$  using an observation-driven Poisson regression model with a latent Markov process, and to study the impact of covariates on its variation. Our estimators allow a close monitor and dynamic update on the knowledge of  $R_t$  whenever new data were available and allow a forecasting of future  $R_t$  under different conditions to provide guidance for policy making. We applied the method to a national dataset with more than 800 counties and 5 million cases in United States. We studied the effects of social distancing, temperature, and county-level characteristics on disease transmission, including percentage of elderly population, population density, and percentage of population with diabetes. Our analysis indicated that of these factors, implementation of social distancing has been the most significant in reducing transmission. We also found population density and percentage of elders were positively associated with  $R_t$ , while temperature had limited impact on reducing SARS-CoV-2 transmission. Based on the estimated effects, we then forecasted the number of new cases counties could experience in the future if different policies were implemented. These results have been used to understand the dynamics of SARS-CoV-2 transmission and provided knowledge to policymakers for making decisions (such as social distancing orders, masking mandates and school reopening strategies).

### Session 89: Current advances in forensic statistics

#### Bayesian Characterizations Of U-processes Used In Pattern Recognition With Application To Forensic Source Identification

♦ *Christopher Saunders<sup>1</sup>, Cami Fuglsby<sup>1</sup>, Danica Ommen<sup>2</sup> and Joann Buscaglia<sup>3</sup>*

<sup>1</sup>South Dakota State University

<sup>2</sup>Iowa State University

<sup>3</sup>FBI Laboratory, Research and Support Unit  
christopher.saunders@sdstate.edu

In forensic science, a typical interpretation task is a common-but-unknown-source identification, where an examiner must summarize and present the evidential value associated with two sets of objects relative to two propositions. The first proposition is that the two sets of objects are two simple random samples from the same, unknown source in a given population of sources; the second proposition is that the two sets of objects are two simple random samples each drawn from two different but unknown sources in a given population of sources. Typically, there is not a natural feature space for which modern statistical techniques can be applied to the non-nested models for model selection. In this presentation, a score function has been developed that maps the trace samples from their measured feature space to the real number line. The resulting score for two trace samples can be used as a measure of the atypicality of matching samples, which will be applied to a receiver operating characteristic (ROC) curve and in a score-based likelihood ratio function. The application of this rule leads to a natural U-process of degree two for assessing the evidence. In this work, we will characterize the U-process and demonstrate how to write a class of approximately admissible decision rules in terms of the U-process. Combining the asymptotic representation of this U-process with an ABC like algorithm, we can then provide summary statistics with Bayes factor-like properties for the selection between the two propositions. We will illustrate this method with an application based on microformometry of aluminum powders, which are common components of IEDs.

#### A Method of Forensic Evidence Interpretation using Error Rates

♦ *Danica Ommen<sup>1</sup>, Larry Tang<sup>2</sup> and Christopher Saunders<sup>3</sup>*

<sup>1</sup>Iowa State University

<sup>2</sup>University of Central Florida

<sup>3</sup>South Dakota State University  
dmommen@iastate.edu

Recent recommendations concerning the use of error rates in forensic science have started to switch the focus of evidence interpretation methods away from the subjective Bayesian approach long advocated by the research community. Some of the concerns related to error rates of forensic science methods are mentioned in the Congressionally mandated 2009 National Academy of Science (NAS) report entitled "Strengthening Forensic Science in the United States: A Path Forward" and the 2016 President's Council of Advisors on Science and Technology (PCAST) report entitled "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods." In response to these recommendations, as well as the success of the black box and white box studies in latent print analysis, a large number of other forensic disciplines have proposed similar studies. These types of studies report an average error rate across a population of examiners for a given set of tasks related to identification of source problems. Although this is not the intention, these studies are used to justify the conclusions that an examiner has made in a specific case. As statisticians focused on the identification of specific source problems, it is our view that it is unclear what these studies imply about a specific source identification problem. In this presentation, we will work within the classical paradigm for evidence interpretation based on conditional match probabilities, a type of forensic error rate. We will then relate these to the error rates common in individuality studies and propose

a paradigm for evidence interpretation based on forensic error rates.

### Univariate Likelihood Ratio Estimation via Mixture of Beta Distributions

♦*Martin Slawski and He Qi*

George Mason University  
mslawsk3@gmu.edu

Likelihood ratios play an important role in forensic evidence interpretation. In this talk, we discuss non-parametric estimation of a univariate likelihood ratio. It is shown that the problem can be formulated as estimating a density on the unit interval, which is then estimated via a mixture of beta approach aka Bernstein polynomials. Unlike kernel density estimation, the approach is effectively free of tuning parameters. Extensions to incorporate certain shape constraints such as likelihood ratio ordering and covariate information are discussed as well.

### Order-Constrained ROC Regression with Application to Facial Recognition

♦*Xiaochen Zhu<sup>1</sup>, Martin Slawski<sup>2</sup> and Larry Tang<sup>3</sup>*

<sup>1</sup>George Mason University

<sup>2</sup>George Mason University

<sup>3</sup>University of Central Florida

xzhu11@masonlive.gmu.edu

The receiver operating characteristic (ROC) curve is widely used to assess discriminative accuracy of two groups based on a continuous score. In a variety of applications, the distributions of such scores across the two groups exhibit a stochastic ordering. Specific examples include calibrated biomarkers in medical diagnostics or the output of matching algorithms in biometric recognition. Incorporating stochastic ordering as an additional constraint into estimation can improve statistical efficiency. In this article, we consider modeling of ROC curves using both the order constraint and covariates associated with each score given that the latter (e.g., demographic characteristics of the underlying subjects) often have a substantial impact on discriminative accuracy. The proposed method is based on the indirect ROC regression approach using a location-scale model, and quadratic optimization is used to implement the order constraint. The statistical properties of the proposed order-constrained least squares estimator are studied. Based on the theoretical results developed herein, we deduce that the proposed estimator can achieve substantial reductions in mean squared error relative to its unconstrained counterpart. Simulation studies corroborate the superior performance of the proposed approach. Its practical usefulness is demonstrated in an application to face recognition data from the "Good, Bad, and Ugly" face challenge, a domain in which accounting for covariates has hardly been studied. Supplementary materials for this article are available online.

## Session 90: Statistical and Machine Learning models on EHR and Insurance Claim databases

### Big Data to answer Big Questions: Experience with Aneurysmal SAH

*Vahed Mafoury<sup>1</sup>, ♦Ashraf Yaseen<sup>1</sup> and George Williams<sup>2</sup>*

<sup>1</sup>UTHSPA

<sup>2</sup>McGovern Medical School  
george.w.williams@uth.tmc.edu

The presenter will discuss aneurysmal subarachnoid hemorrhage and its pathogenesis. Presentation will include statistical and analytic methods associated with a project evaluating factors which predict mortality in this patient population.

### Statistics and Machine Learning Methods for EHR Data: From Data Extraction to Data Analytics/Predictions

*Ashraf Yaseen*

The University of Texas Health Science Center at Houston  
ashraf.yaseen@uth.tmc.edu

This short course will provide an overview and present details of electronic health record (EHR) data extraction, cleaning, processing and analytics for scientific discoveries. The use of EHR data is becoming more prevalent for research purpose and deriving real-world evidence for decision or policy-making. However, analysis of this type of data has many unique complications due to how they are collected, processed, missing data issues, and types of questions that can be answered. This proposed short course covers many important topics related to using EHR data for research and scientific discoveries that include data extraction, cleaning, processing, making inference, and predictions based on many years of practical experience of instructors and their collaborators in the EHR Working Group at the University of Texas Health Science Center at Houston (UTHealth). Statistical and machine learning approaches will also be presented for EHR data extraction, cleaning and analysis. Additionally, since research projects for EHR Big Data are being conducted in large multidisciplinary research groups, the approaches for multiple-project management are necessary and will be also covered in this course

### Inferring Comorbidity Networks from EHR Data

♦*Xi Luo<sup>1</sup>, Gen Zhu<sup>2</sup> and Hulin Wu<sup>2</sup>*

<sup>1</sup>The University of Texas Health Science Center at Houston

<sup>2</sup>University of Texas Health Science Center at Houston  
xi.luo@uth.tmc.edu

Comorbidity networks derived from electronic health records (EHR) has gained more and more attention and can provide insightful information on disease progression and disease management. Pairwise correlation approaches are the most frequently used tools to construct comorbidity networks. Recently, by considering multiple diseases into one model, multivariate methods are applied to characterize the disease relationships from EHR. However, either pairwise methods or existing multivariate methods fail to incorporate the temporal information well. In this paper, we propose an autoregressive model to build the binary disease network from EHR. The model can fully utilize the temporal information and disease status at the encounter level for each patient. Another challenge in comorbidity network analysis is the lack of golden standards for the disease network to compare different methods. A novel simulation study is conducted to compare our method with the pairwise methods and a multivariate method. Our method demonstrates better performance in terms of recovering the true disease network. We also implement the auto-logistic model to a real EHR data set about drug overdose. Symptoms involving the digestive system and chronic pain are found to be the most influential diseases among drug overdose patients.

### Disease Influence Factor and Directed Disease Comorbidity Networks Derived from Big Longitudinal Health Care Data

*Vahed Maroufy*

The University of Texas Health Science Center at Houston  
vahed.maroufy@uth.tmc.edu

The analysis of disease correlations, comorbidities, and progressions serves as a principle to predict the future status of patients in precision medicine. Longitudinal comorbidity networks on population-wide data has not been systematically attempted. In this paper, we studied disease progression patterns using the data from an electronic health records database, Cerner Health Facts, covering



one-sixth of the population in the United States. We used the complete diagnosis records and converted 15 years of health records on 49.5 million people into three temporal comorbidity networks. We also defined disease influence factor (DIF) on the comorbidity networks to quantify the temporal effect of comorbidities at disease level. DIF brought a new insight into quantifying disease severities and provided a novel way to identify fundamental diagnoses in disease progression. For validation and comparison purposes, we applied our methodology to an insurance claim database from Blue Cross Blue Shield of Texas (BCBSTX). Our analyses revealed that complications of pregnancy and diseases of the circulatory system were among the most influential disease categories. We also identified type 2 diabetes and essential kidney diseases among the influential diseases. In addition, although BCBSTX contains a limited population compared to Cerner, the comparison confirmed 73-89% of disease pairs identified in Cerner, and returned the correlation range 0.66-0.80, between the DIFs obtained from the two databases across the three temporal networks.

### Session 91: Utilization of Historical Control Data for Clinical Development

#### To borrow or not to borrow? Determining when historical borrowing has value in a clinical trial

*Kert Viele*

Berry Consultants  
kert@berryconsultants.net

"Those who ignore history are condemned to repeat it". This is the real question of historical borrowing. When historical information is on point and relevant, borrowing can increase estimation precision, reduce type I error, increase power, and lower sample size of clinical trials. When it isn't, we encounter biases, inflated type I error and reduced power. We usually can't determine whether borrowing will be beneficial in advance with certainty, but it is vital we correctly anticipate the possible similarity of our historical and current data, and make decisions on the amount of borrow (aggressively, modestly, not at all) on the basis of that distribution. If we correctly identify our risks, we can choose long run borrowing strategies which benefit the population of trials.

#### Use of Pseudo Controls in Clinical Development

*Larry Shen*

Clinical Informatics-WuXi Apptec  
larry.shen@wuxiapptec.com

In this presentation, I use a few examples to discuss the need for using pseudo or historical controls in clinical trials as part of clinical development programs. I will share some real experiences and practical considerations. There are situations when uncontrolled studies can contribute to a successful drug development program.

#### Explore the use of matching method to supplement clinical trials with historical control data

*Xiang Zhang*

CSL Behring  
xiang.zhang@cslbehring.com

Aggregate historical data (e.g., data from literature) have been utilized in supplementing clinical trials to help decision making and various statistical methods were developed for this utilization (e.g., power priors, robust meta-analytical priors). Recently, patient-level historical control data are becoming more available through the cross-industry data sharing collaboration such as TransCelerate Placebo and Standard of Care Initiative. In this talk, we will explore

the use of matching methods to supplement concurrent clinical trials with patient-level historical controls and simulations are conducted to evaluate the performance of different matching methods.

### Session 92: Statistical development for single-cell RNA-Seq data in biomedical studies

#### scDesign2: a statistical simulator that recapitulates gene correlations for benchmarking scRNA-seq data analysis

♦*Tianyi Sun, Wei Vivian Li and Jingyi Jessica Li*

University of California, Los Angeles  
jli@stat.ucla.edu

In single-cell RNA sequencing (scRNA-seq) experimental design and computational method development, critical questions include how to decide the optimal number of cells to sequence and how to fairly benchmark computational methods. To address these questions, here we developed scDesign2, a flexible statistical simulator that generates synthetic scRNA-seq data mimicking real data from multiple scRNA-seq platforms. A unique feature of scDesign2 is its ability to recapitulate gene-gene correlations, which could not be captured by any existing scRNA-seq simulators. Thanks to the realistic nature of its synthetic data, scDesign2 is a useful tool for single-cell researchers in designing experiments, developing computational methods, and choosing appropriate methods for specific data analysis needs.

#### Integrative differential expression and gene set enrichment analysis for scRNA-seq studies

*Ying Ma<sup>1</sup>, Shiquan Sun<sup>1</sup>, Mengjie Chen<sup>2</sup> and ♦Xiang Zhou<sup>1</sup>*

<sup>1</sup>University of Michigan

<sup>2</sup>University of Chicago  
xzhousph@umich.edu

Differential expression (DE) analysis and gene set enrichment (GSE) analysis are commonly applied in single cell RNA sequencing (scRNA-seq) studies. Here, we develop an integrative and scalable computational method, iDEA, to perform joint DE and GSE analysis through a hierarchical Bayesian framework. By integrating DE and GSE analyses, iDEA can improve the power and consistency of DE analysis and the accuracy of GSE analysis. Importantly, iDEA uses only DE summary statistics as input, enabling effective data modeling through complementing and pairing with various existing DE methods. We illustrate the benefits of iDEA with extensive simulations. We also apply iDEA to analyze three scRNA-seq data sets, where iDEA achieves up to five-fold power gain over existing GSE methods and up to 64% power gain over existing DE methods. The power gain brought by iDEA allows us to identify many pathways that would not be identified by existing approaches in these data.

#### Imputation methods for scRNA-seq data

♦*Wei Vivian Li<sup>1</sup> and Jingyi Jessica Li<sup>2</sup>*

<sup>1</sup>The State University of New Jersey

<sup>2</sup>University of California, Los Angeles  
vivian.li@rutgers.edu

ScRNA-seq data analysis is complicated by excess zero counts, the so-called dropouts due to low amounts of mRNA sequenced within individual cells. We will discuss the imputation problem of scRNA-seq data and summarize available methods. In particular, we introduce scImpute, a statistical method to accurately and robustly impute the dropouts in scRNA-seq data. scImpute automatically identifies likely dropouts, and only perform imputation on these values without introducing new biases to the rest data. Evaluation based on

both simulated and real human and mouse scRNA-seq data suggests that scImpute is an effective tool to recover transcriptome dynamics masked by dropouts. scImpute is shown to identify likely dropouts, enhance the clustering of cell subpopulations, improve the accuracy of differential expression analysis, and aid the study of gene expression dynamics.

#### **RZiMM-scRNA: A regularized zero-inflated mixture model framework for single-cell RNA-seq data**

♦*Xinlei Mi and Jianhua Hu*

Columbia University  
xm2231@cumc.columbia.edu

Applications of single-cell RNA sequencing in various biomedical research areas have been blooming. This new technology provides unprecedented opportunities to study disease heterogeneity at the cellular level. However, unique characteristics of scRNA-seq data, including large dimensionality, high dropout rates, and possibly batch effects, bring great difficulty into the analysis of such data. Not appropriately addressing these issues obstructs true scientific discovery. Herein, we propose a unified Regularized Zero-inflated Mixture Model framework designed for scRNA-seq data (RZiMM-scRNA) to simultaneously detect cell subgroups and identify gene differential expression based on a developed importance score, accounting for both dropouts and batch effects. We conduct extensive simulation studies and perform a real data study on glioma to demonstrate the promise of RZiMM-scRNA in comparison to several popular methods.

### **Session 93: Machine Learning and Real World Data**

#### **New development of statistical machine learning methods with applications to large NHLBI longitudinal studies**

♦*Colin Wu<sup>1</sup>, Xiaoyang Ma<sup>1</sup> and Xin Tian<sup>2</sup>*

<sup>1</sup>National Heart, Lung and Blood Institute

<sup>2</sup>National Heart, Lung and Blood Institute  
wuc@nhlbi.nih.gov

Major epidemiological studies at the National Heart, Lung and Blood Institute (NHLBI) are almost exclusively longitudinal studies with long-term follow-up observations of risk factors (e.g., socioeconomic factor, behavioral status and biomarkers), sub-clinical disease variables (e.g., MRI of the heart, coronary computed tomography, carotid ultrasound and ECG exams) and disease events (e.g., cardiovascular disease, heart failure and mortality). Objectives of these studies include: (a) identifying important risk factors and sub-clinical disease variables that are useful for the prediction of disease events; (b) selecting useful “time-trends” and “landmarks” of influential time-varying risk factors and sub-clinical diseases to predict disease events; (c) developing flexible statistical models for time-varying or dynamic disease prediction and decision making. Traditional statistical methods for flexible longitudinal analysis, such as nonparametric regression models, fall short of Objectives (a) and (b) because they require a small set of prespecified longitudinal predictors. Existing methods of statistical machine learning, such as regularized regression models, fall short of Objectives (b) and (c) because they lack the ability to correctly taking the time-varying nature of the potential predictors and outcomes into consideration. We present a series of recently developed statistical machine learning (ML) methods for correlated and time-varying (i.e., dynamic) feature selection and present their applications to a number of long-term longitudinal studies at NHLBI. We show that these time-varying ML methods are capable of producing biomedically interpretable event prediction models with “functional” and/or

“landmark” longitudinal predictors. In particular, our applications to NHLBI epidemiological studies suggest that the effects of cardiovascular disease risks indeed change with time (e.g., age) and should be treated as dynamic functions of appropriate time variables.

#### **A Practical Guideline for Factor Analysis of Binary/Ordinal Data**

♦*Wen Wan<sup>1</sup>, Vivian Li<sup>2</sup>, Erin Hoefling<sup>3</sup>, Dave Faldmo<sup>3</sup>, Rosy Chang Weir<sup>2</sup> and Marshall Chin<sup>1</sup>*

<sup>1</sup>University of Chicago

<sup>2</sup>Association of Asian Pacific Community Health Organizations

<sup>3</sup>Siouxland Community Health Center  
wenw@uchicago.edu

Applying factor analysis to binary and/or ordinal data seems not difficult by simply treating binary/ordinal as continuous variables. However, there is no a clear practical guideline on how to apply the exploratory factor analysis method for binary/ordinal data to find reasonable latent variables and score them for further analyses. This study aims to develop such a practical guideline through a pragmatic example on the analysis of a PRAPARE social determinants of health (SDH) dataset. Using a cross-sectional observational design, the Siouxland Community Health Center collected the SDH data from their general adult population (total  $n = 11,773$ ) including 716 individuals with diabetes only, 2,388 with hypertension only, 1,477 with both, and 7,192 with neither diabetes/hypertension. We applied the exploratory factor analysis (EFA) method to the 22 SDH variables, which are binary or ordinal. We developed a practical guideline to select appropriate variables as input for an EFA. Then we used the confirmatory factor analysis (CFA) method to evaluate and verify the factor constructs found by EFA. We also applied the ‘sum scores’ approximate method to obtain scores for each latent variable for further association analysis with clinical outcomes including uncontrolled diabetes/hypertension

#### **Dynamic Risk Prediction Using Survival Tree Ensembles**

♦*Yifei Sun<sup>1</sup>, Sy Han Chiou<sup>2</sup>, Colin Wu<sup>3</sup>, Meghan McGarry<sup>4</sup> and Chiung-Yu Huang<sup>5</sup>*

<sup>1</sup>Columbia University

<sup>2</sup>University of Texas at Dallas

<sup>3</sup>National Heart, Lung, and Blood Institute

<sup>4</sup>University of California San Francisco

<sup>5</sup>University of California San Francisco  
ys3072@cumc.columbia.edu

With the availability of massive amounts of data from electronic health records and registry databases, incorporating time-varying patient information to improve risk prediction has attracted great attention. To exploit the growing amount of predictor information over time, we develop a unified framework for landmark prediction using survival tree ensembles, where an updated prediction can be performed when new information becomes available. Compared to the conventional landmark prediction, our framework enjoys great flexibility in that the landmark times can be subject-specific and triggered by an intermediate clinical event. Moreover, the nonparametric approach circumvents the thorny issue in model incompatibility at different landmark times. When both the longitudinal predictors and the outcome event time are subject to right censoring, existing tree-based approaches cannot be directly applied. To tackle the analytical challenges, we consider a risk-set-based ensemble procedure by averaging martingale estimating equations from individual trees. Extensive simulation studies are conducted to evaluate the performance of our methods. The methods are applied to the Cystic Fibrosis Patient Registry (CFPR) data to perform dynamic

prediction of lung disease in cystic fibrosis patients and to identify important prognosis factors.

### **Harnessing Real-World Data for Regulatory Use and Applying Innovative Applications**

♦ *Kelly Zou<sup>1</sup>, Jim Li<sup>1</sup>, Joseph Imperato<sup>1</sup>, Chandrashekhara Potkar<sup>1</sup>, Nikuj Sethi<sup>2</sup>, Jon Edwards<sup>3</sup> and Amrit Ray<sup>2</sup>*

<sup>1</sup>Viatrix

<sup>2</sup>Pfizer

<sup>3</sup>Envision Pharma Group  
kelly.zou@viatrix.com

A vast quantity of real-world data (RWD) are available to healthcare researchers. Such data come from diverse sources such as electronic health records, insurance claims and billing activity, product and disease registries, medical devices used in the home, and applications on mobile devices. The analysis of RWD produces real-world evidence (RWE), which is clinical evidence that provides information about usage and potential benefits or risks of a drug. This presentation focuses on how regulatory authorities are using RWE. The main challenges in harnessing RWD include collating and analyzing numerous disparate types or categories of available information including both structured and unstructured data. Although the use of artificial intelligence to capture, amalgamate, standardize, and analyze RWD is still evolving, it has the potential to support the increased use of RWE to improve healthcare. (Keywords: real-world data, real-world evidence, regulatory, artificial intelligence, robotic process automation.)

### **Session 94: Recent advances in statistical genomics, genetics and EHR data**

#### **WEVar: a novel statistical learning framework for predicting noncoding regulatory variants**

*Li Chen*

Indiana University School of Medicine  
chen61@iu.edu

Understanding the functional consequence of noncoding variants is of great interest. Though genome-wide association studies (GWAS) or quantitative trait locus (QTL) analyses have identified variants associated with traits or molecular phenotypes, most of them are located in the noncoding regions, making the identification of causal variants a particular challenge. Existing computational approaches developed for prioritizing noncoding variants produce inconsistent and even conflicting results. To address these challenges, we propose a novel statistical learning framework, which directly integrates the precomputed functional scores from representative scoring methods. It will maximize the usage of integrated methods by automatically learning the relative contribution of each method and produce an ensemble score as the final prediction. The framework consists of two modes. The first “context-free” mode is trained using curated causal regulatory variants from a wide range of context and is applicable to predict noncoding variants of unknown and diverse context. The second “context-dependent” mode further improves the prediction when the training and testing variants are from the same context. By evaluating the framework via both simulation and empirical studies, we demonstrate that it outperforms integrated scoring methods and the ensemble score successfully prioritizes experimentally validated regulatory variants in multiple risk loci.

#### **A Novel Approach on Multiple-Traits Genetic Association Tests for Flexible Pleiotropy Structures**

*Han Hao*

University of North Texas  
han.hao@unt.edu

Increasing empirical evidence shows the existence of pleiotropy, where genetic variants influence multiple phenotypes related to complex diseases such as glaucoma, hypertension, autism spectrum disorder, major depressive disorder, and schizophrenia. There are two different types of pleiotropy: causal pleiotropy, where genetic variants directly affect multiple phenotypes simultaneously; and mediated pleiotropy, where genetic variants affect certain phenotypes through the mediation of other phenotypes and demographic covariates. Although there are a number of existing multiple traits association tests, few tests can deal with both causal and mediated pleiotropy. We propose a novel multiple-traits genetic association test framework which is flexible for various pleiotropy structures by selecting mediators adaptively. This approach will not only increase the statistical power by aggregating multiple weak effects, but also improve our understanding of the disease etiology.

#### **Fused Landmark Approach for Dynamic Risk Prediction with Application to Electronic Health Record Data**

♦ *Jiehuan Sun<sup>1</sup>, Katherine Liao<sup>2</sup> and Tianxi Cai<sup>3</sup>*

<sup>1</sup>University of Illinois at Chicago

<sup>2</sup>Harvard Medical School

<sup>3</sup>Harvard T.H. Chan School of Public Health  
jiehuan.sun@gmail.com

Many studies have been conducted on dynamic risk prediction models, which take advantage of the data property of longitudinal variables and can be used to predict the risk of an event for a subject while the risk can be updated as more information is collected for the subject. However, most existing dynamic risk prediction models can only deal with one or a few longitudinal variables. It is essential to develop a dynamic risk prediction model that can deal with high-dimensional longitudinal data, as the data nowadays usually have a large number of longitudinal variables, such as the longitudinal electronic health record data. To that end, we propose a novel fused landmark method in this article. Our proposed method builds upon the traditional landmark method and uses fused lasso penalty to allow for information borrowing across different landmark time points and variable selection. Through extensive simulation studies, we show that our proposed method outperforms traditional landmark methods in both prediction and variable selection. We then apply our method to a dataset of patients with type 2 diabetes and find that our method has better predictive performance than other landmark methods and identifies a few variables that have interesting time-dependent effects.

### **Session 95: Real World Evidence Study in Healthcare**

#### **Deep Learning for Analyzing Electronic Health Records**

*Fei Wang*

Weill Cornell Medicine  
few2001@med.cornell.edu

Deep learning (DL) models have achieved big success in many applications including computer vision, speech analysis and natural language processing. In healthcare, DL has also demonstrated strong potentials in medical image analysis and clinical natural language processing. Recently, researchers have also been actively trying to develop effective DL models for analyzing longitudinal electronic health records (EHR). In this talk I will present some examples of this line of research, point out the pitfalls and challenges, as well as future directions.

### Statistical Anomaly Detection in Dynamic Brain Networks

♦ *Dorcas Ofori-Boateng*<sup>1</sup>, *Ivor Cribben*<sup>2</sup> and *Yulia Gel*<sup>3</sup>

<sup>1</sup>Portland State University

<sup>2</sup>University of Alberta

<sup>3</sup>University of Texas at Dallas  
doforib2@pdx.edu

Detecting change points and anomalies in dynamic network structures has become increasingly popular across many disciplines and, especially, in neuroscience, where an important objective is the reconstruction of the dynamic mechanism underlying brain region interactions. Indeed, analysis of changes in human brain connectivity has profound clinical implications because it constitutes an important step towards more accurate diagnostics of many neuropsychiatric diseases and better understanding of disease stage progression. However, most current statistical methods for detecting anomalies suffer from the following limitation: network snapshots at different time points are assumed to be independent. To circumvent this limitation, we propose a distribution-free framework for anomaly detection in dynamic networks. First, we present each network snapshot of the data as a linear object and find its respective univariate characterization via local and global network topological summaries. Second, we adopt a change point detection method for (weakly) dependent time series based on efficient scores and enhance the finite sample properties of change point method by approximating the asymptotic distribution of the test statistic using the sieve bootstrap. We illustrate the utility of the proposed approach for time-evolving graphs by applying it to a task based functional magnetic resonance imaging (fMRI) experiment where we have a good sense of where the change points occur. We also apply our new approach to a resting-state fMRI experiment, where mental activity is unconstrained. We find that our new method delivers impressively accurate and realistic results in terms of identifying locations of true change points compared to the results reported by competing approaches. The new method has the potential to unveil the time-varying cognitive states of both controls and subjects with neuropsychiatric diseases such as Alzheimer's, dementia, autism, and schizophrenia in order to develop new understandings of these diseases. Furthermore, using the new approach, we can consider whole brain dynamics, which promises to offer deeper insight into the large-scale characterizations of functional architecture of the whole brain.

### Disease screening for a personality disorder using Electronic Health Records (EHR) data

♦ *Nan Shao*<sup>1</sup>, *Marianne Goodman*<sup>2</sup>, *Chengxi Zang*<sup>3</sup> and *Vikas Mohan Sharma*<sup>1</sup>

<sup>1</sup>Boehringer Ingelheim

<sup>2</sup>Icahn School of Medicine at Mount Sinai; James J Peters VA Medical Center

<sup>3</sup>Weill Cornell Medicine, Cornell University  
nan.shao@boehringer-ingelheim.com

Borderline personality disorder (BoPD) is one of the most common personality disorders marked by an ongoing pattern of varying moods, self-image, and behavioral issues. We have developed a machine-learning algorithm to automatically screen likely BoPD patients who are currently not formally diagnosed with BoPD, using electronic health record (EHR) data. In this talk, we will discuss motivation, data, methodology, challenges, and model performance.

### Estimation of Individualized Treatment Rules Using a Covariate-Specific Treatment Effect Curve

♦ *Wenchuan Guo*<sup>1</sup>, *Xiao-Hua Zhou*<sup>2</sup> and *Shujie Ma*<sup>3</sup>

<sup>1</sup>Bristol-Myers Squibb

<sup>2</sup>Peking University

<sup>3</sup>University of California Riverside  
wenchuan.guo@bms.com

With a large number of baseline covariates, we propose a new semi-parametric modeling strategy for heterogeneous treatment effect estimation and individualized treatment selection. We model the covariate-specific treatment effect curve as an unknown function of a weighted linear combination of all baseline covariates. The weight or the coefficient for each covariate is estimated by fitting a sparse semi-parametric logistic single-index coefficient model. The CSTE curve is then estimated by a spline-backfitted kernel procedure, which enables us to further construct a simultaneous confidence band for the CSTE curve under a desired confidence level. Based on the confidence band, we find the subgroups of patients who benefit from the treatment, so that we can make individualized treatment selection.

### Session 96: Bayesian Additive Regression Tree: Theory, Computation, and Application

#### Multidimensional Monotonicity Discovery with MBART

♦ *Robert McCulloch*<sup>1</sup> and *Edward George*<sup>2</sup>

<sup>1</sup>Arizona State University

<sup>2</sup>University of Pennsylvania  
remccul1@asu.edu

For the discovery of a regression relationship between  $y$  and  $x$ , a vector of  $p$  potential predictors, the flexible nonparametric nature of BART (Bayesian Additive Regression Trees) allows for a much richer set of possibilities than restrictive parametric approaches. To exploit the potential monotonicity of the predictors, we introduce mBART, a constrained version of BART that incorporates monotonicity with a multivariate basis of monotone trees, thereby avoiding the further confines of a full parametric form. Using mBART to estimate such effects yields (i) function estimates that are smoother and more interpretable, (ii) better out-of-sample predictive performance and (iii) less post-data uncertainty. By using mBART to simultaneously estimate both the increasing and the decreasing regions of a predictor, mBART opens up a new approach to the discovery and estimation of the decomposition of a function into its monotone components.

#### Bayesian Decision Tree Ensembles in Fully Nonparametric Problems

♦ *Antonio Linero*<sup>1</sup>, *Yinpu Li*<sup>2</sup> and *Jared Murray*<sup>1</sup>

<sup>1</sup>University of Texas at Austin

<sup>2</sup>Florida State University  
antonio.linero@austin.utexas.edu

We introduce several extensions of BART to fully-nonparametric problems which are based on modulating an underlying Poisson process. We illustrate how our framework can be used to solve conditional distribution and nonparametric survival analysis problems. The BART models we propose are often simpler to use than more common Bayesian nonparametric approaches while enjoying similar benefits in terms of allowing for the prior to be "centered" on a desired parametric/semiparametric model. Our approach allows for simple Gibbs sampling algorithms which use the familiar Bayesian backfitting approach commonly used to fit BART models. Taking advantage of the strong theoretical properties of certain BART priors, we are able to establish posterior concentration at near-minimax optimal rates for these problems, adaptively over a large class of

function spaces. We illustrate our methodology on simulated and benchmark datasets.

### **Causal Inference and Sensitivity Analysis for Unmeasured Confounding in Observational Data with Multiple Treatments and a Binary Outcome**

♦ *Liangyuan Hu<sup>1</sup>, Chenyang Gu<sup>2</sup>, Michael Lopez<sup>3</sup>, Jiayi Ji<sup>1</sup> and Juan Wisnivesky<sup>1</sup>*

<sup>1</sup>Icahn School of Medicine

<sup>2</sup>Analysis Group, Inc.

<sup>3</sup>Skidmore College

liangyuan.hu@mountsinai.org

There is a dearth of robust methods to estimate the causal effects of multiple treatments when the outcome is binary. And sensitivity analysis methods for unmeasured confounding in this context are sparse. We investigate the operating characteristics of Bayesian Additive Regression Trees (BART) for causal inference in such settings, and compare BART to several approaches that have been proposed for continuous outcomes, including inverse probability of treatment weighting (IPTW), targeted maximum likelihood estimator (TMLE), vector matching and regression adjustment. We then develop a Monte Carlo sensitivity analysis approach for the complex multiple treatment settings using BART for flexibly modeling the response surfaces. We first derive the general bias form introduced by unmeasured confounding (UMC), with emphasis on theoretical properties uniquely relevant to multiple treatments. We then propose methods to encode the impact of UMC on the potential outcomes and adjust the estimates of causal effects in which the presumed UMC is removed. Expansive simulations provide empirical evidence for the validity of our methods and gain insights into sensitivity analysis strategies in the multiple treatment setting. A comprehensive sensitivity analysis of the SEER-Medicare data elucidates the comparative causal effects of three surgical approaches among early stage non-small cell lung cancer patients in respect of four postoperative patient outcomes.

### **Stochastic tree ensembles for regularized nonlinear regression**

♦ *Jingyu He<sup>1</sup> and P. Richard Hahn<sup>2</sup>*

<sup>1</sup>University of Chicago

<sup>2</sup>Arizona State University

jingyu.he@chicagobooth.edu

This paper develops a novel stochastic tree ensemble method for nonlinear regression, which we refer to as XBART, short for Accelerated Bayesian Additive Regression Trees. By combining regularization and stochastic search strategies from Bayesian modeling with computationally efficient techniques from recursive partitioning approaches, the new method attains state-of-the-art performance: in many settings it is both faster and more accurate than the widely-used XGBoost algorithm. Via careful simulation studies, we demonstrate that our new approach provides accurate pointwise estimates of the mean function and does so faster than popular alternatives, such as BART, XGBoost and neural networks (using Keras). We also prove a number of basic theoretical results about the new algorithm, including consistency of the single tree version of the model and stationarity of the Markov chain produced by the ensemble version. Furthermore, we demonstrate that initializing standard Bayesian additive regression trees Markov chain Monte Carlo (MCMC) at XBART-fitted trees considerably improves credible interval coverage and reduces total run-time.

### **Session 97: New Methods for Missing Data in Public Health Studies**

#### **An augmented survival analysis method for mis-measured and interval censored outcomes**

♦ *Chongliang Luo, Rebecca Hubbard and Yong Chen*

University of Pennsylvania

chongliang.luo@penmedicine.upenn.edu

Survival analyses with interval censored outcomes are commonly seen in clinical and epidemiological studies, where the patients are only examined periodically and the event or failure of interest is known only to occur within a certain interval. Often the examination procedure is also subject to mis-measurement error. For example, in Electronic health records (EHR) based association study, phenotypes of patients are derived from a high-throughput phenotyping algorithm, whereas chart reviews (deemed as a gold standard for the true phenotype) are available only for a small subset of patients. We provide a method that jointly use the error-prone outcomes of all patients and true outcomes of a validation sample to achieve bias reduction and also efficiency improvement. We presented simulation and real data examples to compare our method with other available methods.

#### **Deep Learning for Time-to-event Outcomes**

♦ *Jon Steingrimsson, Samantha Morrison and Constantine Gatsonis*

Brown University

jon.steingrimsson@brown.edu

Deep learning is a class of algorithms that uses multiple layers to create a risk prediction model. The layers involve an unknown weight vector that is estimated by minimizing a loss function. We extend the deep learning algorithms to handle censoring by utilizing semi-parametric efficiency theory for missing data to develop loss functions that can be calculated in the presence of censoring. We discuss properties of these loss functions and practical issues related to implementation of the deep learning algorithms. We furthermore discuss extensions to convolutional neural networks for imaging analysis. The performance of the resulting algorithms is evaluated through simulation studies and by analyzing data on breast cancer patients.

#### **A New Bayesian Joint Model for Mixed Types of Longitudinal Data in the Presence of Different Missing Data Patterns with Applications to HIV Prevention Trials**

*Jing Wu*

University of Rhode Island

jing\_wu@uri.edu

In longitudinal clinical trials, it is common that mixed types of outcomes are collected on the same subject over time. It is also routinely encountered that all outcomes may be subject to substantial missing values due to dropout and intermittent missingness. Additionally, the missing data patterns of the mixed types of outcomes are usually the same for dropout while different for intermittent missingness. In this paper, a sequential multinomial model is adopted for dropout and subsequently, a new joint conditional model is constructed for intermittent missingness of mixed types of outcomes. The new model captures the complex structure of missingness and incorporates dropout and different intermittent missingness simultaneously. Two types of outcomes (binary and count) are considered in this paper. A mixed-effects probit regression model and a zero-inflated Poisson mixed-effects regression model are assumed for the longitudinal binary and count response data, respectively. We further show that the joint posterior distribution is improper if uniform priors are specified for the regression coefficients

under the proposed model. An efficient Gibbs sampling algorithm is developed using a hierarchical centering technique. A modified logarithm of the pseudomarginal likelihood (LPML) and a new concordance measure criterion are used to compare the models under different missing data mechanisms. An extensive simulation study is conducted to investigate the empirical performance of the proposed methods, and the methods are further illustrated using real data from an HIV prevention clinical trial.

#### **Bayesian Modeling and Inference for Item Response Model with Nonignorable Missing Data**

Jing Wu<sup>1</sup>, ♦Zhihua Ma<sup>2</sup> and Ming-Hui Chen<sup>3</sup>

<sup>1</sup>University of Rhode Island

<sup>2</sup>Shenzhen University

<sup>3</sup>University of Connecticut

mazh1993@outlook.com

Not-reached (dropout) and omitted (intermittent missingness) items are often inevitable in timed tests where answers are not required. The missingness of the item response may be related to the subject's latent characteristics, the difficulty, or even the unobserved response of the item. To fully understand the underlying results of the testing, we must handle the missing data appropriately. In this article, we propose a new missing data mechanism, which jointly studies the not-reached and omitted behaviors for the multilevel item response theory (IRT) model. This proposed methodology is illustrated using real data from the Program for International Student Assessment (PISA) 2015 study. A modified deviance information criterion (DIC) is developed to assess model fit. Extensive simulations are conducted to further illustrate the generality of the proposed model, and show that our proposed model compares favorably with another competing model.

#### **Session 98: Keynote speech**

##### **Towards a Blend of Statistics and Microeconomics**

Michael I. Jordan

University of California, Berkeley

jordan@cs.berkeley.edu

Statistical decisions are often given meaning in the context of other decisions, particularly when there are scarce resources to be shared. Managing such sharing is one of the classical goals of microeconomics, and it is given new relevance in the modern setting of large, human-focused datasets, and in data-analytic contexts such as classifiers and recommendation systems. I'll discuss several recent projects that aim to explore this interface, including the study of exploration-exploitation trade-offs for bandits that compete over a scarce resource, notions of local optimality in nonconvex-nonconcave minimax optimization and how such notions relate to stochastic gradient methods, the use of Langevin-based algorithms for Thompson sampling, and multi-agent learning based on online gradient descent.

#### **Session 99: Innovative Statistical and data science methods for clinical trial studies**

##### **Phase I/II seamless designs in oncology trials**

Inna Perevozskaya

GSK

inna.x.perevozskaya@gsk.com

Phase I dose-escalation in oncology has seen a dramatic uptake of innovative design usage over the past decade. Many companies are

adopting a model-based approach versus traditional 3+3 designs, but even these improved methods cannot dramatically increase precision of MTD finding due to their limited sample size and restrictive dose exploration. There is added on complexity of evaluating multiple drugs (combinations), multiple grades of toxicity or simultaneous assessment of target engagement and toxicity in a single study. In other words, Phase I oncology trials of today are very different from small, MTD-focused studies of a single drug with dichotomous endpoint of the past. They are larger, often include substantial expansion cohorts, and their objectives are more complex than simple MTD determination as a single dose to take forward. Such goals are better addressed by incorporating a seamless Phase 2-like extension into initial dose-escalation phase and using combined data along with quantitative decision making to improve the probability of success on further development.

##### **Time to Endpoint Maturation Framework and Application**

♦Li Wang<sup>1</sup>, Mengjia Yu<sup>2</sup> and Hongtao Zhang<sup>3</sup>

<sup>1</sup>AbbVie

<sup>2</sup>UIUC

<sup>3</sup>Celgene

wangleelee@gmail.com

In clinical trials, an accurate prediction of certain key milestone dates will have significant impact on trial planning, monitoring and execution. Time to Endpoint Maturation (TTEM) framework which is an extension of Time to Event (TTE) is developed to model and predict any milestone date that is defined by number of events of interest. It is an integration of time to accrual and time to event (subject to censoring) when accrued. In event driven trials, both parts need advanced statistical modeling while in non-event driven trials, only time to accrual matters. New methodologies on both time to accrual and time to event are developed to achieve a better prediction accuracy comparing to existing methods in commercial software

##### **Response adaptive randomization designs and implementation in clinical trials**

Lanju Zhang

AbbVie

lanju.zhang@abbvie.com

Adaptive randomization has been proposed for both ethical and/or efficiency purposes. Instead of using the same randomization ratio throughout a clinical trial, the randomization ratio is updated based on accumulative in-trial data. There is a rich literature of this topic with two major approaches: Bayesian adaptive randomization and a frequentist approach. However, adaptive randomization is rarely used in industry-sponsored trials. In this presentation, we will review new developments in response adaptive randomization design, the current regulatory guidance, its niche application areas, practical considerations, and our experiences with regulatory agency.

##### **Precision medicine: subgroup identification in clinical trials**

♦Lei Liu and Chamila Perera

Washington University in St. Louis

lei.liu@wustl.edu

One of the major topics in precision medicine is subgroup identification when the treatment is efficacious or effective for some subjects but shows a negligible or even detrimental effect for others. Recently proposed recursive partitioning methods play a major role in identifying such post hoc subgroups. In this paper we are interested in identifying patient subgroups responding differently to the topical treatment on the risk of primary open angle glaucoma (POAG) in Phase I of the Ocular Hypertension Treatment Study (OHTS I). We

applied an interaction tree (IT) procedure which used recursively partitioning methods to identify participant subgroups with different interaction effects of covariates at baseline and treatment on the risk of developing POAG. Due to a substantial portion of missing values for some variables, we first conducted multiple imputation using software IVEware1 accounting for correlation between two eyes, resulting in five imputed data sets. We applied the IT procedure to each of the imputed datasets to identify subgroups with heterogeneous treatment effects. Though overall the treatment effect is significant (hazard ratio (HR)=0.40,  $p < 0.001$ )<sup>2</sup>, the treatment effect is heterogeneous. Our methods identified four subgroups based on demographic and clinical factors at baseline. The two subgroups with 274 subjects (HR=0.94,  $p=0.89$ ) and 311 subjects (HR=1.32,  $p=0.41$ ) had no statistically significant difference between medication and control in terms of POAG risk. The remaining two subgroups with 909 subjects (HR=0.17,  $p<0.001$ ) and 142 subjects (HR=0.31,  $p=0.02$ ) had a significantly decreased risk of POAG in the medication group as compared to the control group.

### Session 100: New methods in semiparametric inferences for analyzing real world data

#### Empirical Likelihood for varying coefficient Geo Models

Shuoyang Wang

Auburn University  
szw0100@auburn.edu

In this work, we introduce a varying coefficient geo model for spatial data distributed over complex domains. The univariate components and the geographical component in the model are approximated via univariate polynomial splines and bivariate penalized splines over triangulation, respectively. We also propose test procedures based on the empirical likelihood with bias-corrected estimating equations to conduct both pointwise and simultaneous inferences. The asymptotic distributions of the test statistics are derived under the null and local alternative hypotheses. Simulation studies and real data analyses are conducted to demonstrate the methods proposed.

#### Semi-parametric multinomial logistic regression for multivariate point pattern data

Kristian Hesselund<sup>1</sup>, ♦Ganggang Xu<sup>2</sup>, Yongtao Guan<sup>2</sup> and Rasmus Waagepetersen<sup>1</sup>

<sup>1</sup>Aalborg University

<sup>2</sup>University of Miami  
gangxu@bus.miami.edu

We propose a new method for analysis of multivariate point pattern data observed in a heterogeneous environment and with complex intensity functions. We suggest semi-parametric models for the intensity functions that depend on an unspecified factor common to all types of points. This is for example well suited for analyzing spatial covariate effects on events such as street crime activities that occur in a complex urban environment. A multinomial conditional composite likelihood function is introduced for estimation of intensity function regression parameters and the asymptotic joint distribution of the resulting estimators is derived under mild conditions. Crucially, the asymptotic covariance matrix depends on the cross pair correlation functions of the multivariate point process. To make valid statistical inference without restrictive assumptions, we construct consistent non-parametric estimators for cross pair correlation function ratios. Finally, we construct standardized residual plots and predictive probability plots to validate and to visualize the

findings of the model. The effectiveness of the proposed methodology is demonstrated through extensive simulation studies and an application to analyzing the effects of socio-economic and demographic variables on occurrences of street crimes in Washington DC.

### Session 101: Recent development in Semiparametric regression analysis

#### Set regression with application in subgroup analysis

Ao Yuan

Georgetown University  
ay312@georgetown.edu

Regression models are commonly used in statistical analysis. It is the conditional mean of the response given covariates  $\mu(bx) = E(Y|bX = bx)$ . However, in some practical problems, the interest is the conditional mean of the response given where the covariates belong to some set  $A$ . Notably, in precision medicine and subgroup analysis in clinical trials, we aim to identify subjects who benefit the most from the treatment, namely, identify an optimal set in the covariate space which manifests treatment favoritism if a subject's covariates fall in this set and the subject is classified to the favorable treatment subgroup. Existing methods for subgroup analysis achieve this indirectly by using classical regression. This motivates us to develop a new type of regression: it set-regression, defined as  $\mu(A) = E(Y|bX \text{ in } A)$  which directly addresses the subgroup analysis problem. This extends the classical regression model but also improves recursive partitioning and support vector machine approaches, and is particularly suitable for objectives involving optimization of the regression over sets, such as subgroup analysis.

#### Semiparametric inference for marginal and association parameters in the distribution of bivariate event times data

Dongdong Li<sup>1</sup>, Joan Hu<sup>2</sup> and ♦Rui Wang<sup>1</sup>

<sup>1</sup>Harvard Pilgrim Health Care Institute and Harvard Medical School

<sup>2</sup>Simon Fraser University  
rwang@hsph.harvard.edu

Evaluating association between two event times can be challenging when observations on the event times are subject to informative censoring due to a terminating event such as death. Moreover, methods for assessing covariate effects on association are lacking in this context. We propose an approach that simultaneously models the marginal and association parameters and their dependence on covariates accounting for informative censoring. This is done by modeling the joint distribution of the two event times and the informative censoring time in nested copula functions and incorporating flexible functional forms of covariate effects in both the marginal and association parameters. We develop an easy-to-implement pseudolikelihood-based inference procedure, derive the asymptotic properties of the resulting estimators, and conduct simulation studies to examine the finite sample performance of the proposed approach. For illustration, we apply the proposed approach to a breast cancer study to characterize the association between cardiovascular disease and relapse/second cancer in the presence of death as a semi-competing risk.

#### Joint Penalized Spline Modeling of Multivariate Longitudinal Data

♦Lihui Zhao<sup>1</sup>, Tom Chen<sup>2</sup>, Vladimir Novitsky<sup>2</sup> and Rui Wang<sup>2</sup>

<sup>1</sup>Northwestern University

<sup>2</sup>Harvard University

lihui.zhao@northwestern.edu

Motivated by the need to jointly model the longitudinal trajectories of HIV viral load levels and CD4 counts during the primary infection stage, we propose a joint penalized spline modeling approach that can be used to model the repeated measurements from multiple biomarkers of various types (eg, continuous, binary) simultaneously. This approach allows for flexible trajectories for each marker, accounts for potentially time-varying correlation between markers, and is robust to misspecification of knots. Despite its advantages, the application of multivariate penalized spline models, especially when biomarkers may be of different data types, has been limited in part due to its seemingly complexity in implementation. To overcome this, we describe a procedure that transforms the multivariate setting to the univariate one, and then makes use of the generalized linear mixed effect model representation of a penalized spline model to facilitate its implementation with standard statistical software. We performed simulation studies to evaluate the validity and efficiency through joint modeling of correlated biomarkers measured longitudinally compared to the univariate modeling approach. We applied this modeling approach to longitudinal HIV-1 RNA load and CD4 count data from Southern African cohorts to estimate features of the joint distributions such as the correlation and the proportion of subjects with high viral load levels and high CD4 cell counts over time.

#### Estimation and testing for extended partially linear single-index models

Zijuan Chen and ♦Suojin Wang

Texas A&M University  
sjwang@stat.tamu.edu

In partially linear single-index models, there are two different covariate matrices in the model for the linear part and non-linear part. All covariate information needs to be divided into two parts before the model can be fitted. In contrast, in the extended partial linear single index models, all the covariate variables are included in one matrix, which is contained in both the linear part and non-linear part of the model. We propose local smoothing estimators for the model parameters and unknown function, whose asymptotic properties are demonstrated. We also employ the LASSO penalty to obtain penalized estimators with consistency and oracle property in order to carry out estimation and variable selection simultaneously. Finally, we develop a linear hypothesis test for the model parameters. Simulation studies are presented that support our analytic results. In addition, a real data analysis is provided for illustration.

#### Session 102: Stochastic gradient Monte Carlo for big data statistics

##### Extended Stochastic Gradient MCMC for Large-Scale Bayesian Variable Selection

♦Qifan Song, Yan Sun, Mao Ye and Faming Liang

Purdue University  
qfsong@purdue.edu

Stochastic gradient Markov chain Monte Carlo (MCMC) algorithms have received much attention in Bayesian computing for big data problems, but they are only applicable to a small class of problems for which the parameter space has a fixed dimension and the log-posterior density is differentiable with respect to the parameters. This paper proposes an extended stochastic gradient MCMC algorithm which, by introducing appropriate latent variables, can be applied to more general large-scale Bayesian computing prob-

lems, such as those involving dimension jumping and missing data. Numerical studies show that the proposed algorithm is highly scalable and much more efficient than traditional MCMC algorithms. The proposed algorithms have much alleviated the pain of Bayesian methods in big data computing.

##### Optimal-Transport Bayesian Sampling

Changyou Chen

University at Buffalo  
changyou@buffalo.edu

Deep learning has achieved great success in recent years. One aspect overlooked by traditional deep-learning methods is uncertainty modeling, which can be very important in certain applications such as medical image classification and reinforcement learning. A standard way for uncertainty modeling is by adopting Bayesian inference. In this talk, I will share some of our recent work on scalable Bayesian inference by sampling, called optimal-transport sampling, motivated from the optimal-transport theory. Our framework formulates Bayesian sampling as optimizing a set of particles, overcoming some intrinsic issues of standard Bayesian sampling algorithms such as sampling efficiency and algorithm scalability. I will also describe how our sampling framework be applied to uncertainty and repulsive attention modeling in state-of-the-art natural-language-processing models.

##### Stochastic Gradient MCMC for Sequential Decision Making

Yan Ma

UC San Diego  
yianma.ucsd@gmail.com

In this talk, I'll discuss how to design stochastic gradient MCMC algorithms for the task of sequential decision making. Importantly, we need to ensure that optimal regret is achieved with a constant computational budget. That requires us to have increasingly accurate estimation with growing data set, under constant number of iterations and computation per iteration. I will present an approximate stochastic gradient Langevin dynamics that achieves this goal.

##### An Adaptively Weighted Stochastic Gradient MCMC Algorithm for Global Optimization in Deep Learning

Wei Deng, Guang Lin and ♦Faming Liang

Purdue University  
fmliang@purdue.edu

We propose an adaptively weighted stochastic gradient Langevin dynamics (AWSGLD) for Bayesian learning in big data statistics. The proposed algorithm is scalable, which automatically adjusts the target distribution, flattening the high energy region and protruding the low energy region, such that global optimization can be greatly facilitated. Theoretically, we establish the convergence of the AWSGLD algorithm and provide an upper bound for its hitting time to the optimal set for a wide class of non-convex functions. The AWSGLD algorithm has a much shorter hitting time than stochastic gradient Langevin dynamics (SGLD). AWSGLD is tested on multiple benchmark datasets including CIFAR10 and CIFAR100. The numerical results indicate its superiority over the existing state-of-the-art algorithms in training deep neural networks.

#### Session 103: Statistical and AI inferences based on DNA and protein sequences

##### Stairway Plot 2: demographic history inference with folded SNP frequency spectra

♦Xiaoming Liu<sup>1</sup> and Yun-Xin Fu<sup>2</sup>

<sup>1</sup>University of South Florida



<sup>2</sup>The University of Texas Health Science Center at Houston  
xiaomingliu@usf.edu

Inferring the demographic histories of populations has wide applications in population, ecological, and conservation genomics. We present Stairway Plot 2, a cross-platform program package for this task using SNP frequency spectra (SFSs). It is based on a nonparametric method with the capability of handling folded SFSs (i.e., the ancestral alleles of the SNPs are unknown) of thousands of samples produced with genotyping-by-sequencing technologies; therefore, it is particularly suitable for nonmodel organisms.

#### **Origin of protein collective motions: a case study on serine protease proteinase K**

*Shu-Qun Liu*

Yunnan University  
shuqunliu@gmail.com

To probe the origin of the protein collective motions, multiple long molecular dynamics (MD) simulations on serine protease proteinase K with the solute and solvent coupled to different temperatures (either 300 or 180 K) was conducted. Comparative analyses demonstrate that the internal flexibility and mobility of proteinase K are strongly dependent on the solvent temperatures but weakly on the temperatures of the protein itself. The constructed free energy landscapes (FELs) at the high solvent temperatures exhibit a more rugged surface, broader spanning range, and higher minimum free energy level than do those at the low solvent temperatures. Comparison between the dynamic hydrogen bond (HB) numbers reveals that the high solvent temperatures intensify the competitive HB interactions between water molecules and protein surface atoms, and this in turn exacerbates the competitive HB interactions between protein internal atoms, thus enhancing the conformational flexibility and facilitating the collective motions of the protein. A refined FEL model was proposed to explain the role of the solvent mobility in facilitating the cascade amplification of microscopic motions of atoms and atomic groups into the global collective motions of the protein.

#### **Statistical inferring the clonal and subclonal architecture of cancer genomes**

*Yupeng Cun*

Chinese Academy of Sciences  
yp.cun@outlook.com

The genomes of cancer cells are constantly reshaped during pathogenesis. This evolutionary process leads to the emergence of subclonal populations, which can limit therapeutic interventions by the emergence of drug-resistance mutations. Data derived from massively parallel sequencing can be used to infer these subclonal populations from tumor-specific point mutations. The accurate determination of copy number changes and tumor impurity is an indispensable requirement to reliably infer these subclonal populations by mutational clustering. This protocol describes a copy number analysis method together with a novel mutational clustering approach. The method is called Sclust. In a series of simulations and comparisons with alternative methods, we showed that Sclust accurately determines copy number states and subclonal populations. Performance tests showed that the entire method is computationally extremely efficient. In particular, copy number analysis and mutational clustering takes less than 10 minutes

#### **Supervised learning for analyzing large-scale genome-wide DNA polymorphism data**

*Haipeng Li*

Chinese Academy of Sciences  
lihaipeng@picb.ac.cn

Supervised learning has been extensively applied in many fields; Alpha-GO and autopilot might be two of the most well-known cases. However, its application in population and evolutionary genetics is still in childhood. Recently, we introduced the boosting, a supervised learning approach, to identify positive Darwinian selection in natural populations and estimate recombination rate along the human genome. We further analyzed the genome-wide DNA polymorphism data from nearly 10,000 human individuals (UK10K) and obtained a fine-scale genetic map for humans. The number of identified autosomal recombination hotspots is about 2.93-14.25 times less than that previously identified in human populations, indicating that the variance of estimated recombination rate may be underestimated when identifying recombination hotspots, especially population-specific human recombination hotspots. These results indicate that supervised learning approaches, together with deep learning and reinforced learning, could play essential roles when analyzing large-scale genome-wide DNA polymorphism data.

#### **Session 104: Advanced topics in causal inference**

##### **Variable Selection for Causal Mediation Analysis Using LASSO-based Methods**

*Zhaoxin Ye<sup>1</sup>, Yeying Zhu<sup>1</sup> and Donna Coffman<sup>2</sup>*

<sup>1</sup>University of Waterloo

<sup>2</sup>Temple University  
yeying.zhu@uwaterloo.ca

Unbiased causal mediation effect estimates can be obtained from marginal structural models using inverse probability weighting with appropriate weights. In order to compute weights, treatment and mediator propensity score models need to be fitted first. If the covariates are high-dimensional, parsimonious propensity score models can be developed by regularization methods including LASSO and its variants. Furthermore, in a mediation setup, efficient direct or indirect effect estimators can be obtained by using outcome-adaptive LASSO to select variables for propensity score models by incorporating the outcome information. A simulation study is conducted to assess how different regularization methods can affect the performance of estimated natural direct and indirect effect odds ratios. Our simulation results suggest that bias reduction can be achieved with sufficient covariate balancing and propensity score models regularized by outcome-adaptive LASSO can be used to improve the efficiency of the natural direct effect estimator. The regularization methods are then applied to MIMIC-III database, an ICU database developed by MIT.

##### **On regression approach to propensity score analysis**

*Liang Li*

The University of Texas MD Anderson Cancer Center  
lli115@mdanderson.org

In propensity score analysis, the frequently used regression adjustment involves regressing the outcome on the estimated propensity score and treatment indicator. This approach can be highly efficient when model assumptions are valid, but can lead to biased results when the assumptions are violated. We extend the simple regression adjustment to a varying coefficient regression model that allows for nonlinear association between outcome and propensity score. We discuss its connection with some propensity score matching and weighting methods, and show that the proposed analytical framework can shed light on the intrinsic connection among some mainstream propensity score approaches (stratification, regression, kernel matching, and inverse probability weighting) and handle com-

monly used causal estimands. We derive analytic point and variance estimators that properly take into account the sampling variability in the estimated propensity score. Extensive simulations show that the proposed approach possesses desired finite sample properties and demonstrates competitive performance in comparison with other methods estimating the same causal estimand. The proposed methodology is illustrated with a study on right heart catheterization.

### **SSc/SSc-ILD Patient Journey: Data-driven Disease Trajectories in EHR/Claims Databases**

*Yahui Tian*

Boehringer Ingelheim

yahui.tian@boehringer-ingelheim.com

Population wide analysis of disease correlations and patterns of progression is a promising path towards precision medicine, because a vital prerequisite for choosing the right treatment for the right person is the estimation of disease development from the current state. Electronic health records (EHR) and health insurance administrative databases provide rich and inexpensive sources of research information to facilitate evidence-based medicine, for example to promote clinical decision support and analyze clinical outcomes under specific conditions. We propose to use sequential data mining methods in high-dimensional real world databases, the goal is to identify

significantly temporal patterns of disease progression in Systemic Scleroderma (SSc) and SSc with Interstitial Lung Disease (SSc-ILD), results could help in comprehensive disease understanding and early disease diagnosis.

### **Bayesian Additive Regression Trees for Causal Inference**

*Dai Feng*

AbbVie Inc.

dai.feng@abbvie.com

Various methods have been proposed to identify causal effects. Many approaches require fitting two models: one for the assignment mechanism and one for the response surface. A different strategy focusing on flexible modeling of the response surface only using a Bayesian non-parametric model: Bayesian Additive Regression Trees (BART), has been proposed. The advantages of using BART for causal inference include accommodation of different types of outcomes (continuous, categorical and time-to-event), requirement of less assumptions on model construction, handling of a large number of predictors, coherent estimate of uncertainty and identification of heterogeneous treatment effects. In this talk, I will introduce how to use BART for causal inference, share results from a simulation study for evaluation of the performance of different approaches, illustrate the advantages of BART, and outline recent and relevant methodology advancements.

# Index of Authors

- A.salem, L, 43, 107  
 Alemdjrodo, K, 36, **39**, 82, **94**  
 Allen, G, 46, 119  
 Altieri, N, 45, 117  
 Amorim, GGC, 38, 90  
 Armstrong, GT, 28, 52  
 Avram, D, 40, 97
- Bai, R, 41, 103  
 Bakoyannis, G, 39, 94  
 Balocchi, C, 41, 103  
 Barrientos, A, **32**, **65**  
 Barter, RL, 45, 117  
 Bauer, C, **34**, **72**  
 Beauchamp, M, 45, 115  
 Becker, J, 35, 77  
 Beckman,R, **31**, **63**  
 Bedi, T, **38**, **87**  
 Bedoui, A, 36, 82  
 Bhadra, A, **44**, **113**  
 Bi, X, **37**, **84**  
 Bian, Y, **33**, **69**  
 Bickel, P, 41, 102  
 Bliznyuk, N, **40**, **99**  
 Boerwinkle, E, 30, 43, 59, 111  
 Bowden, J, 33, 68  
 Bradic, J, 46, 119  
 Bradley, J, **40**, **99**  
 Branson, M, 41, 101  
 Brantley, H, 41, 102  
 Buelvas, B, 94  
 Bunn, V, 33, 43, 69, 108  
 Buscaglia, J, 46, 121
- Cabral, M, 35, 75  
 Cai, B, **39**, **95**  
 Cai, J, 30, 57  
 Cai, T, 47, 125  
 Cantu, E, 36, 80  
 Cao, C, 34, 72  
 Cao, S, 29, 53  
 Cao, X, **40**, **99**  
 Casadebaig, M, 41, 101  
 Chakraborty, S, 32, 66  
 Chan, GMA, 28, 50  
 Chan, KW, 39, 95  
 Chan, NH, 40, 98
- Chang, C, 38, 40, 44, 90, 98, 113  
 Chang, V, **34**, **73**  
 Chang, Y, 34, 73  
 Chatterjee, N, 34, 73  
 Chen, B, **36**, **80**  
 Chen, C, 38, **48**, 91, **130**  
 Chen, H, **41**, **46**, **100**, **120**  
 Chen, J, 29, **35**, **42**, 45, 53, **76**, **104**, 115  
 Chen, Josh, 39, 96  
 Chen, K, **40**, 44, **98**, 115  
 Chen, L, 31, 36, **41**, **45**, **47**, 60, 80, **102**, **116**, **125**  
 Chen, M, 34, **36**, 38, 47, 48, 72, **79**, 87, 123, 128  
 Chen, Q, **35**, **76**  
 Chen, S, **30**, **40**, 41, **57**, **97**, 100  
 Chen, T, 48, 129  
 Chen, W, 44, 115  
 Chen, Y, **30**, **31**, 31, **32**, 33, 38, 47, **60**, 61, **62**, **67**, 68, 91, 127  
 Chen, Z, 48, 130  
 Cheng, C, 29, 55  
 Cheng, F, 39, 92  
 Cheng, Q, 28, 51  
 Cheng, Y, **35**, **36**, 39, 41, **77**, **82**, 91, 101  
 Chi, E, 38, 39, **41**, 88, 89, 94, **102**  
 Chin, M, 47, 124  
 Chinchilli, VM, 44, 114  
 Chiou, SH, 47, 124  
 Chiu, C, 31, **43**, 61, **110**  
 Choi, J, 33, 71  
 Christensen, J, **33**, **69**  
 Chu, C, 34, 71  
 Chu, H, 32, 67  
 Coad, DS, 40, 99  
 Coffman, D, 49, 131  
 Colditz, G, 45, 116  
 Conklin, H, 29, 55  
 Constantine, F, 44, 111  
 Cook, J, **42**, **106**
- Cook, R, 29, 54  
 Crawford, F, 37, 86  
 Cribben, I, 47, 126  
 Cui, Y, **32**, 36, **43**, **66**, 79, **107**  
 Cui, Z, 41, 101  
 Cun, Y, **49**, **131**
- Daniels, M, **36**, **80**  
 Dasarathy, G, 46, 119  
 Datta, J, **41**, **100**  
 Demeulemeester,J, 29, 53  
 Deng, H, 43, 111  
 Deng, M, **39**, **92**  
 Deng, Q, 34, 40, 74, 98  
 Deng, W, 48, 130  
 Deng, X, 41, 100  
 Deng, Y, 28, 50  
 Deshpande, S, **41**, **103**  
 Diao, L, **29**, **54**  
 Dickey, E, 35, 75  
 Ding, J, 37, 85  
 Ding, X, 32, 65  
 Ding, Y, **33**, 44, **71**, 115  
 Dmitrienko, A, 33, 69  
 Do, K, 35, 75  
 Dong, G, 34, 73  
 Dong, M, 43, 107  
 Dong, X, 38, 90  
 Du, P, 31, **37**, 37, 61, 82, **83**  
 Duan, C, **39**, **92**  
 Duan, J, 31, 60  
 Duerr, R, 44, 115  
 Duncan, J, 45, 117  
 Dunson, D, 32, 65  
 Duren, Y, 35, 76  
 Dwivedi, R, 45, 117
- Edwards, J, 43, 47, 107, 125  
 Engle, H, **38**, **87**  
 Erlendsdottir, M, **37**, **86**
- Faldmo, D, 47, 124  
 Fan, R, 43, 110  
 Fan, X, 38, 90  
 Fan, Z, **38**, **91**  
 Fang, S, 30, 59  
 Fang, Y, 40, 42, **45**, 98, 106, **118**
- Fang, Z, 33, 71  
 Faries, D, 41, 101  
 Felizzi, F, 30, 56  
 Feng, D, **49**, **132**  
 Feng, L, 37, 84  
 Feng, R, **36**, **80**  
 Feng, Y, 31, 45, 62, 116  
 Feng, Z, 37, 86  
 Fernandezmorales, E, **38**, **90**  
 Fine, J, 35, 77  
 Fisher, J, 36, 79  
 Foster, R, 34, 72  
 Fougeres, A, 40, 96  
 Franz, T, 34, 74  
 Fu, H, 44, 113  
 Fu, Y, 49, 130  
 Fuglsby, C, 46, 121
- Gagnon-Bartsch, J, **28**, **53**  
 Gajjar, A, 29, 55  
 Gamalo-Siebers, M, 44, 111  
 Gambino, G, 34, 72  
 Gao, F, **45**, **116**  
 Gao, L, 30, 59  
 Gao, Z, **37**, **82**  
 Garcia, T, 30, 57  
 Gatsonis, C, 48, 127  
 Ge, Q, 30, **43**, 59, **110**  
 Gel, Y, 47, 126  
 Geng, P, **34**, **75**  
 George, E, 47, 126  
 George, S, 43, 108  
 Gilbert, P, 29, 35, 56, 77  
 Gleason, K, 31, 60  
 Goldstein, I, 42, 106  
 Goldwasser, M, 37, 85  
 Gong, J, 41, 101  
 Goodman, M, 47, 126  
 Goren, A, 42, 106  
 Griffin, M, 35, 76  
 Gu, C, 47, 127  
 Gu, E, 30, 56  
 Guan, Y, 39, 48, 95, 129  
 Guinness, J, 41, 102  
 Guo, H, 41, 100  
 Guo, J, 44, 112  
 Guo, S, 29, 30, 53, 59  
 Guo, W, **47**, **126**  
 Guo, X, 30, 31, 58, 62

Guo, Y, **39, 95**  
 Guo, Z, 39, 41, 91, 101  
  
 Ha, MJ, **42, 104**  
 Hahn, PR, 47, 127  
 Han, D, **37, 39, 83, 92**  
 Han, P, **30, 60**  
 Han, X, **38, 89**  
 Hao, H, **47, 125**  
 Harrar, S, 32, 66  
 Harrell, F, 35, 76  
 Hassan, T, **42, 106**  
 He, J, **47, 127**  
 He, L, 31, 64  
 He, W, 34, 42, 45, 74, 106, 118  
 He, X, 31, 60  
 He, Z, **33, 40, 68, 97**  
 He, ZC, **40, 97**  
 Heathfield, A, 42, 106  
 Hemani, G, 33, 68  
 Heng, F, **29, 35, 56, 77**  
 Hesselund, K, 48, 129  
 Hoaglin, D, 34, 73  
 Hoefling, E, 47, 124  
 Hogan, J, 44, 115  
 Hu, G, 32, 65  
 Hu, J, 47, 48, 124, 129  
 Hu, J, **41, 101**  
 Hu, L, **47, 127**  
 Hu, T, **31, 63**  
 Hu, X, 45, 118  
 Hu, Y, 41, 101  
 Hu, Z, 30, **42, 59, 104**  
 Hu,Z, 43, 110  
 Huang, B, 34, 73  
 Huang, C, **39, 47, 92, 124**  
 Huang, H, 40, 44, 98, 115  
 Huang, J, **46, 121**  
 Huang, S, 31, 62  
 Huang, X, **30, 39, 59, 93**  
 Hubbard, R, 47, 127  
 Hunt, G, 28, **31, 53, 63**  
 Huo, S, 44, 113  
 Huo, Z, **40, 97**  
 Hyun, S, 29, 56  
  
 Iaci, R, **36, 81**  
 Ibrahim, J, 36, 45, 79, 118  
 Imperato, J, **42, 43, 47, 107, 107, 125**  
 Ing, C, 40, 98, 99  
 Ionita-Laza, I, 29, 54  
 Ishida, E, **45, 118**  
  
 Jenq, R, 35, 75  
 Jeong, J, 38, 89  
 Ji, J, 36, 47, 80, 127  
 Ji, S, 29, 53  
 Ji, Y, 37, **45, 84, 117**  
 Jia, Y, **38, 89**  
  
 Jiang, D, 36, 81  
 Jiang, S, 33, **38, 68, 88**  
 Jiang, T, **45, 118**  
 Jiang, X, **38, 87**  
 Jiang, Y, 35, **36, 39, 77, 81, 92**  
 Jiao, J, **32, 65**  
 Jin, B, **33, 68**  
 Jin, H, 31, 61  
 Jin, L, **30, 30, 43, 59, 60, 110**  
 Jin, P, **35, 39, 41, 78, 91, 93, 101**  
 Jin, W, 28, **38, 52, 87**  
 Jin, Y, 42, 107  
 Jinnah, A, **35, 78**  
 Johnson, E, **35, 76**  
 Jordan, Michael, 48, 128  
 Joseph, R, 35, 75  
  
 Kadziola, Z, 41, 101  
 Kan, S, 38, 91  
 Kang, D, 33, 70  
 Kang, J, 40, 97  
 Kashlak, A, 31, 61  
 Ke, C, **30, 58**  
 Kersey, J, **35, 77**  
 Kim, I, **42, 104**  
 Kim, J, **33, 70**  
 Kim, M, **36, 81**  
 Kim, R, **30, 57**  
 Klinedinst, B, 31, 62  
 Koh, A, 33, 68  
 Kolassa, J, 44, 114  
 Kolluri, S, 33, 69  
 Kong, D, 30, 60  
 Kong, L, 31, 61  
 Kong, X, **32, 66**  
 Kordzakhia, G, **33, 69**  
 Koslovsky, M, **41, 103**  
 Krischer, J, 41, 101  
 Krishna, A, **35, 75**  
 Kumbier, K, 45, 117  
 Kwon, J, 44, 114  
  
 Laird, G, 31, 63  
 Lawrence, T, 28, 53  
 Lebeau, J, 35, 75  
 Lederer, J, 35, 76  
 Lee, C, 28, 50  
 Lee, J, 30, 37, 59, 86  
 Lee, JJ, 37, 85  
 Lee, K, 40, 99  
 Lee, M, 39, 93  
 Lee, S, **33, 71**  
 Lei, KC, **39, 39, 91, 92**  
 Leng, D, 38, 91  
 Leng, DL, 39, 92  
 Leon Novelo, L, **34, 72**  
 Leung, MF, **39, 95**  
 Levina, E, 37, 84  
  
 Li, C, **28, 39, 51, 93**  
 Li, D, 31, 41, 42, **43, 48, 64, 101, 105, 109, 129**  
 Li, G, **28, 50**  
 Li, H, **49, 131**  
 Li, J, **41, 43, 47, 100, 107, 125**  
 Li, JJ, 46, 47, 123  
 Li, L, **49, 131**  
 Li, M, 28, **29, 35, 38, 41, 43, 45, 51, 53, 75, 88, 101, 108, 115**  
 Li, P, **36, 82**  
 Li, Q, 33, **34, 38, 41, 44, 68, 74, 87, 88, 101, 114**  
 Li, R, 37, 39, **44, 86, 93, 114**  
 Li, S, 30, 43, 59, 110  
 Li, T, **37, 84**  
 Li, V, 47, 124  
 Li, W, **29, 30, 37, 55, 59, 86**  
 Li, WV, 46, **47, 123, 123**  
 Li, X, **31, 31, 45, 61, 62, 117**  
 Li, Y, 28, **29, 29, 32, 33, 36, 40, 42, 47, 52, 54, 55, 56, 66, 67, 80, 97, 105, 126**  
 Li, Z, 38, **39, 41, 43, 44, 91, 91, 101, 103, 111, 114**  
 Liang, F, **48, 48, 130, 130**  
 Liao, J, **45, 45, 119, 119**  
 Liao, K, 47, 125  
 Liao, Q, 34, 44, 74, 112  
 Liebert, R, 42, 106  
 Lin, D, 28, 29, 51, 54, 55  
 Lin, G, 48, 130  
 Lin, H, **33, 38, 71, 88**  
 Lin, J, **33, 39, 43, 44, 69, 93, 108, 111**  
 Lin, L, **32, 67**  
 Lin, R, **33, 37, 42, 70, 86, 105**  
 Lin, W, 30, 59  
 Lin, X, **29, 29, 45, 55, 56, 117**  
 Lin, X, **43, 109**  
 Lin, Xihong, 29, 55  
 Lin,W, 43, 110  
 Linero, A, **47, 126**  
 Lipkovich, I, **41, 101**  
 Liu, D, **32, 38, 67, 91**  
 Liu, F, **45, 45, 119, 119**  
 Liu, J, **28, 31, 37, 45, 45, 51, 61, 83, 116, 116**  
 Liu, L, 45, **48, 116, 128**  
 Liu, M, 29, 35, 39, 55, 78, 93  
 Liu, P, **39, 95**  
  
 Liu, R, 33, **43, 69, 108**  
 Liu, S, 43, **49, 108, 131**  
 Liu, X, 35, **39, 49, 75, 94, 130**  
 Liu, Y, 30, **34, 45, 58, 72, 118**  
 Liu,Y, 30, **31, 59, 63**  
 Loh, JM, 38, 89  
 Long, Q, **41, 102**  
 Loo,PV, 29, 53  
 Lopez, M, 47, 127  
 Lotspeich, SC, **38, 90**  
 Lou, Y, **31, 63**  
 Lu, B, 34, 72  
 Lu, C, 45, 119  
 Lu, Q, **36, 39, 78, 93**  
 Lu, W, 30, 56  
 Lu, Y, 30, 31, 60, 62  
 Lu, Z, **29, 55**  
 Luan, G, **38, 89**  
 Luo, C, **47, 127**  
 Luo, J, 45, 116  
 Luo, W, 38, 91  
 Luo, X, **34, 41, 46, 71, 101, 122**  
  
 Ma, G, 38, 91  
 Ma, S, 47, 126  
 Ma, X, 47, 124  
 Ma, Y, 34, 47, **48, 73, 123, 130**  
 Ma, Z, **48, 128**  
 Mafoury, V, 46, 122  
 Magnotti, J, 45, 115  
 Mai, Q, **30, 58**  
 Mai, Y, 34, 74  
 Manuel, C, 37, 83  
 Maronge, J, 36, 80  
 Maroufy, V, **46, 122**  
 Mcculloch, R, **47, 126**  
 MCGarry, M, 47, 124  
 Meng, L, 40, 97  
 Mercadier, C, 40, 96  
 Merrill, H, 40, 99  
 Mi, X, **47, 124**  
 Miao, H, 37, 83  
 Michail, S, 43, 108  
 Milena, M, 94  
 Miller, M, 35, 75  
 Min, EJ, 41, 102  
 Mishne, G, 38, 88  
 Montierth, MD, 29, 53  
 Moon, C, **36, 82**  
 Morris, J, **37, 38, 84, 88**  
 Morris, JS, 44, 113  
 Morrison, S, 48, 127  
 Moustaki, I, 31, 62  
 Mu, J, **36, 82**  
 Mueller, P, **37, 84**  
 Mueller, S, 30, 57  
 Murad, MH, 32, 67

Murray, J, 47, 126  
Müller, P, 42, 104  
  
Nadon, R, **32, 64**  
Nan, B, **44, 114**  
Netzorg, R, 45, 117  
Ng, H, 38, 91  
Ni, Y, 28, **37, 37, 38, 52, 83,**  
84, 87  
Nian, H, 35, 76  
Ning, J, 37, **39, 86, 93**  
Ning, Y, **46, 120**  
Niu, C, 35, 75  
Nolan, J, **40, 96**  
Novitsky, V, 48, 129  
  
Ofori-Boateng, D, **47, 126**  
Ommen, D, **46, 46, 121, 121**  
Oorschot, J, 40, 96  
  
Paganizani, C, 42, 104  
Page, G, 32, 65  
Palmer, K, 38, 87  
Pan, J, 37, 85  
Pappadopoulos, E, 42, 107  
Park, B, 45, 117  
Park, Y, 40, **43, 98, 108**  
Parmigiani, G, 35, 76  
Patil, P, **35, 76**  
Patra, B, 38, 90  
Paul, L, 29, 54  
Paulon, G, **42, 104**  
Peng, S, **46, 120**  
Perera, C, 48, 128  
Perevozkaya, I, **48, 128**  
Peterson, C, 35, **42, 75, 104**  
Peterson, CB, 44, 115  
Pierce, B, 31, 60  
Pokal, S, 34, 74  
Potkar, C, 47, 125  
Pritchett, Y, 44, 111  
Psioda, M, **42, 45, 106, 118**  
  
Qi, G, **34, 73**  
Qi, H, 46, 122  
Qi, L, 35, 77  
Qi, X, **44, 115**  
Qin, L, **35, 76**  
Qiu, Y, **31, 62**  
Qu, A, 28, 44, 50, 113  
Quan, H, 34, 71  
  
Ramirez, R, 94  
Ramprasad, R, 35, 75  
Rasmussen, E, 33, 69  
Rathouz, P, **36, 80**  
Ratitch, B, 41, 101  
Ray, A, 43, 47, 107, 125  
Ray, D, **34, 73**  
Reddick, W, 29, 55  
Reich, B, **35, 75**  
  
Ren, B, 35, 76  
Ren, J, 43, 108  
Resnick, S, 40, 96  
Reyers, M, **32, 65**  
Robertson, J, 37, 82  
Robison, L, 29, 56  
Rochani, H, 35, 77  
Roychoudhury, S, **34, 73**  
Rubin, L, 28, 38, 52, 87  
Rui, S, 40, 98  
  
Sadinle, M, 30, 60  
Samawi, H, 35, 77  
Sandra, S, **94**  
Sang, P, **31, 61**  
Saunders, C, **46, 46, 121,**  
121  
Schifano, E, 36, 79  
Schipper, M, 28, 53  
Schmidt, M, 32, 65  
Seaberg, E, 35, 77  
Seifu, Y, 34, 73  
Sepanski, J, **39, 93**  
Sethi, N, 47, 125  
Shang, Z, 37, 82, 83  
Shao, N, **47, 126**  
Shao, Y, **31, 64**  
Sharma, VM, 47, 126  
Sharpton, T, 36, 81  
Shaw, PA, 38, 90  
Shen, D, 32, 65  
Shen, JP, 29, 53  
Shen, L, **46, 123**  
Shen, X, 36, 78  
Sheng, W, **36, 81**  
Shepherd, BE, 38, 90  
Shi, J, 38, 46, 90, 121  
Shi, L, **40, 98**  
Shi, Q, 33, 70  
Shi, Y, 35, 75  
Si, T, 42, 107  
Si, Y, **34, 72**  
Siegel, L, **32, 67**  
Singh, C, 45, 117  
Sinha, S, **37, 83**  
Sinks, S, 34, 72  
Slawski, M, **46, 46, 122, 122**  
Small, DS, 33, 68  
Smith, GD, 33, 68  
Song, C, **40, 98**  
Song, F, 28, 50  
Song, J, **34, 73**  
Song, P, **45, 117**  
Song, Q, **48, 130**  
Song, X, **36, 80**  
Song, Y, **39, 44, 94, 112**  
Spence, A, 38, 87  
Spieker, A, **41, 103**  
Starling, J, 41, 103  
Steingrimsson, J, **48, 127**  
Su, Y, **32, 66**  
  
Sun, B, 38, 91  
Sun, D, **39, 39, 95, 95**  
Sun, F, 29, 43, 54, 108  
Sun, J, 29, 39, **40, 45, 47,**  
56, 95, **96, 116,**  
**125**  
Sun, J(, 39, 95  
Sun, P, 34, 72  
Sun, S, 32, **39, 47, 64, 92,**  
123  
Sun, SX, 39, 92  
Sun, T, 33, **46, 71, 123**  
Sun, W, **29, 54**  
Sun, X, **31, 61**  
Sun, Y, 29, **35, 47, 48, 56,**  
**77, 124, 130**  
Sun, Z, 44, 115  
Swartz, T, 32, 65  
  
Takeda, K, **33, 70**  
Talbot, K, 35, 76  
Tan, YS, 45, 117  
Tang, B, 29, 55  
Tang, CY, **30, 58**  
Tang, L, **44, 46, 112, 121,**  
122  
Tang, M, **28, 53**  
Tang, R, 45, 118  
Tang, W, 42, 106  
Tang, X, **28, 31, 40, 50, 61,**  
99  
Tang, T, 45, 117  
Tao, R, **30, 38, 57, 90**  
Taylor-Rodriguez, D, 34, 72  
Teng, Z, **45, 118**  
Thall, P, **28, 43, 52, 108**  
Tian, C, 36, 81  
Tian, H, **45, 118**  
Tian, L, **46**  
Tian, X, 47, 124  
Tian, Y, **49, 132**  
Ting, N, 34, 74  
Tong, X, 36, 39, 78, 93  
Toyoizumi, K, **34, 74**  
Tran, H, 35, 75  
Trasande, L, 39, 93  
Trippa, L, **44, 112**  
Tsai, H, 40, 99  
Tseng, G, **40, 40, 97, 98**  
Tsung, F, 37, 83  
Tu, W, 40, 97  
  
Urbanucci, A, 29, 53  
  
Vannucci, M, 41, 103  
Vehik, K, 41, 101  
Vernon, C, 37, 86  
Viele, K, **46, 123**  
Vinci, G, **46, 119**  
  
Waagepetersen, R, 48, 129  
  
Wager, S, 46, 119  
Wan, F, **45, 116**  
Wan, W, **47, 124**  
Wang, A, 33, 69  
Wang, C, 29, **33, 45, 46, 56,**  
**70, 117, 120**  
Wang, D, 32, **38, 38, 64, 91,**  
91  
Wang, F, **47, 125**  
Wang, G, 31, **38, 42, 62, 91,**  
107  
Wang, H, **36, 42, 45, 78, 79,**  
**106, 118**  
Wang, J, **31, 31, 32, 33, 36,**  
60, **62, 64, 68, 79**  
Wang, JR, 29, 53  
Wang, L, 28, **29, 29, 30, 30,**  
31, **36, 36, 38,**  
**48, 53, 55, 55,**  
56, 59, **60, 62,**  
**80, 81, 90, 128**  
Wang, M, **28, 44, 50, 114**  
Wang, P, 36, 80  
Wang, R, **48, 48, 129, 129**  
Wang, S, 30, 37, **38, 44, 48,**  
60, 83, **89, 112,**  
**113, 129, 130**  
Wang, T, **29, 40, 43, 54, 96,**  
108  
Wang, X, 29, **37, 43, 44, 46,**  
55, **86, 108, 115,**  
120  
Wang, X(, **29, 55**  
Wang, Y, **29, 31, 39, 44, 54,**  
**62, 93, 115**  
Wang, Z, **28, 35, 42, 43, 45,**  
**51, 77, 105, 110,**  
**115**  
Wang, Wltxbf, 29, 53  
Wang, Y, 45, 117  
Wei, S, 39, 92  
Wei, Y, **28, 29, 50, 54**  
Weir, RC, 47, 124  
Weko, C, 41, 102  
Weng, H, **31, 62**  
Weng, RC, **40, 99**  
Willette, A, 31, 62  
Williams, G, 46, 122  
Wisnivesky, J, 47, 127  
Womack, A, **34, 72**  
Wong, C, 38, 91  
Wong, KY, **28, 51**  
Wong, M, 38, 91  
Wu, C, **47, 47, 124, 124**  
Wu, D, 28, 51  
Wu, F, 34, 72  
Wu, H, 38, 39, 41, **46, 46,**  
90, 91, 101, **120,**  
122  
Wu, J, **32, 32, 36, 48, 48, 65,**  
66, 79, **127, 128**

Wu, L, **45, 119**  
 Wu, M, **34, 74**  
 Wu, Q, **38, 40, 88, 97**  
 Wu, W, **30, 31, 43, 45, 58,**  
     **61, 110, 119**  
 Wu, X, **43, 109**  
 Wu, Y, 35, 39, 76, 93  
 Wu, Z, **28, 41, 51, 103**  
  
 Xi, B, **41, 100**  
 Xia, A, 45, 118  
 Xian, J, 37, 83  
 Xiang, D, **37, 83**  
 Xiao, G, 33, 38, 44, 68, 87,  
     114  
 Xie, J, **34, 74**  
 Xie, Y, 38, 40, 87, 97  
 Xin, H, 44, 115  
 Xing, X, 37, 83  
 Xiong, C, 45, 116  
 Xiong, M, **30, 30, 43, 58,**  
     59, 110, 111  
 Xiu, L, 45, 118  
 Xu, C, **33, 69**  
 Xu, G, **48, 129**  
 Xu, H, 29, **31, 54, 63**  
 Xu, J, **37, 45, 85, 118**  
 Xu, M, **29, 54**  
 Xu, R, **40, 97**  
 Xu, T, 30, 37, **43, 43, 58, 85,**  
     **110, 111**  
 Xu, Y, **28, 34, 38, 44, 44, 52,**  
     74, 87, 111, **115**  
 Xu, Z, 44, 115  
 Xue, F, **44, 113**  
 Xue, S, 41, 102  
 Xue, Y, 32, 65  
  
 Yan, F, 31, 42, 64, 105  
 Yan, J, 32, 65  
 Yan, L, 42, 107  
  
 Yan, M, 36, 80  
 Yang, B, 37, 85  
 Yang, C, 29, 54  
 Yang, F, **31, 60**  
 Yang, H, **32, 43, 65, 107**  
 Yang, LL, 42, 107  
 Yang, P, 29, 53  
 Yao, B, 33, 68  
 Yaseen, A, 38, **46, 90, 122**  
 Ye, C, **37, 85**  
 Ye, J, **41, 42, 105**  
 Ye, M, 48, 130  
 Ye, Z, 49, 131  
 Yi, B, **34, 71**  
 Yi, G, 36, 80  
 Yi, N, 35, 78  
 Yiannoutsos, C, 39, 94  
 Yin, J, 35, 77  
 Yin, J, **33, 70**  
 Yin, X, 30, 36, 58, 81  
 Ying, Z, 31, 61  
 Yu, B, 45, 117  
 Yu, G, 29, 56  
 Yu, L, 31, 63  
 Yu, M, 48, 128  
 Yu, S, **30, 59**  
 Yu, T, **40, 99**  
 Yu, W, 42, 107  
 Yu, Z, 40, 42, 97, 105  
 Yuan, A, 36, **48, 80, 129**  
 Yuan, Q, 36, 81  
 Yuan, Y, **28, 31, 31, 37, 42,**  
     42, 43, **44, 50,**  
     **64, 64, 85, 86,**  
     **105, 105, 108,**  
     **111**  
 Yung, G, **44, 44, 111, 112**  
  
 Zang, C, 47, 126  
 Zeleniuch-Jacquotte, A, 35,  
     78  
  
 Zeng, D, 28, 51  
 Zeng, J, 30, 58  
 Zhan, X, **33, 38, 68, 88**  
 Zhang, B, 31, **43, 61, 110**  
 Zhang, C, **38, 44, 87, 114**  
 Zhang, G, **30, 60**  
 Zhang, H, 32, 37, **45, 48, 64,**  
     84, **117, 128**  
 Zhang, J, **30, 36, 39, 44, 56,**  
     79, 93, 95, **114**  
 Zhang, K, 43, 111  
 Zhang, L, **35, 41, 42, 48, 75,**  
     **100, 105, 128**  
 Zhang, M, **38, 38, 87, 88**  
 Zhang, N, 38, 91  
 Zhang, NR, 33, 68  
 Zhang, P, **28, 53**  
 Zhang, T, **38, 38, 90, 91**  
 Zhang, V, 37, 85  
 Zhang, W, 33, **37, 44, 69,**  
     **85, 112**  
 Zhang, X, **28, 29, 30, 32, 35,**  
     35, 38, **41, 43,**  
     45, **46, 52, 55,**  
     58, **64, 66, 77,**  
     **78, 91, 102, 108,**  
     115, **123**  
 Zhang, XD, 38, 39, 90, 92  
 Zhang, XHD, 39, 92  
 Zhang, Y, **34, 35, 37, 39, 44,**  
     **74, 76, 85, 94,**  
     113, 115  
 Zhang, Y, **43, 109**  
 Zhang, Z, **32, 42, 66, 105**  
 Zhang, C, 45, 117  
 Zhang, K, 43, 110  
 Zhao, H, 43, 108  
 Zhao, L, **48, 129**  
 Zhao, Q, **32, 33, 45, 65, 68,**  
     118  
 Zhao, S, **35, 77**  
  
 Zhao, X, 29, 53  
 Zhao, Y, **36, 36, 37, 39, 41,**  
     **82, 82, 85, 94,**  
     **102**  
 Zhao, Z, **40, 96**  
 Zhao, Y, 35, 39, 78, 95  
 Zheng, J, 34, 73  
 Zheng, P, 38, 91  
 Zhong, A, 39, 92  
 Zhong, J, 34, 72  
 Zhou, C, **40, 96**  
 Zhou, H, 30, **43, 45, 57, 108,**  
     **115**  
 Zhou, J, **31, 43, 61, 109**  
 Zhou, L, 44, 112  
 Zhou, Q, **30, 38, 57, 88**  
 Zhou, S, 44, 115  
 Zhou, T, 45, 117  
 Zhou, W, **39, 39, 44, 44, 94,**  
     94, **112, 112, 113**  
 Zhou, X, **47, 47, 123, 126**  
 Zhou, Y, 29–31, **34, 37, 42,**  
     42, 55, 60, 64,  
     **74, 86, 105, 106**  
 Zhu, G, 46, 122  
 Zhu, H, **44, 113**  
 Zhu, J, 37, **38, 40, 41, 44,**  
     84, **90, 97, 102,**  
     **111**  
 Zhu, L, **28, 29, 36, 52, 56,**  
     81  
 Zhu, X, **34, 46, 73, 122**  
 Zhu, Y, 35, 38, **46, 49, 76,**  
     90, **119, 131**  
 Zhu, Z, **43, 108**  
 Zhuang, W, 35, 78  
 Zhuo, B, 33, 69  
 Zohner, YE, **38, 88**  
 Zou, F, **43, 107**  
 Zou, K, 42, 43, **47, 107, 125**

# See you in 2021 ICOSA Applied Statistics Symposium

[www.icsa.org](http://www.icsa.org)



International Chinese Statistical Association

泛華統計協會