May 20-23, 2021 (Virtual)

IISA 2021 Conference

Statistics in the Era of Evidence Based Inference

Grateful Acknowledgements:
Support from the United States National Science Foundation
Local arrangements and virtual hosting by University of Illinois Chicago and its School of Public Health
Support from Merck Sharp & Dohme, Kenilworth, NJ, USA.

Website: https://www.intindstat.org/summerConference2021/index

Statistics in the Era of Evidence Based Inference
International Indian Statistical Association
University of Illinois, Chicago
20 - 23 May, 2021

Virtual Conference

This Programme Booklet does not contain live links to the zoom rooms. The links next to the venue of your session will take you to the conference registrants login system. The live zoom links are available in a similar booklet on the other side. More instructions are available on page (iv).

Scientific Programme Committee

Chair: Saonli BASU, University of Minnesota
Co-chair: Sanjay CHAUDHURI, National University of Singapore
Co-chair: Satrajit ROYCHOUDHURY, Pfizer Inc.

Sumanta BASU, Cornell University
Swati BISWAS, University of Texas Dallas
Ivan CHAN, Bristol Myers Squibb
Snigdhansu CHATTERJEE, University of Minnesota
Brandon COOMBES, Mayo Clinic
Tirthankar DASGUPTA, Rutgers University
Abhirup DATTA, Johns Hopkins University
Gauri Sankar DATTA, University of Georgia
Tanujit DEY, Brigham and Women's Hospital
Joyee GHOSH, Universoty of Iowa
Adityanand GUNTUBOYINA, University of California, Berkeley
Donald HEDEKER, University of Chicago
Kshitij KHARE, University of Florida
Hrishikesh KULKARNI, Alexion
Prakash LAUD, Medical College of Wisconsin
Rong LIU, Bristol Myers Squibb
Arnab MAITY, Pfizer Inc.
Tapabrata MAITY, Michigan State University
Abhyuday MANDAL, University of Georgia
Debashis MONDAL, Oregon State University
Pabak MUKHOPADHYAY, Daiichi Sankyo
Saurabh MUKHOPADHYAY, AbbVie Inc
Naveen NARISETTY, University of Illinois at Urbana-Champaign
Debashis PAUL, University of California Davis
Karen Lynn PRICE, Eli Lilly and Co.
Debashree RAY, Johns Hopkins University
Bodhisattva SEN, Columbia University
Shanthi SETHURAMAN, Eli Lilly and Co
Karthik SRIRAM, Indian Institute of Management, Ahmedabad
David VOCK, University of Minnesota
Julian WOLFSON, University of Minnesota
Min YANG, University of Illinois at Chicago
Lin ZHANG, University of Minnesota

Acknowledgements:

University of Illinois School of Public Health:
Wayne GILES, Dean
Ronald HERSHOW
Daniel ALBERT
Frank CERVONE
Olivia DESORMEAUX
Shelley HOOVER
Robin SYLVESTER
and all members of UIC SPH

International Indian Statsitical Association:
Sanjib BASU, University of Illinois at Chicago
Snigdhansu CHATTERJEE, University of Minnesota
Jyotishka DATTA, Virginia Tech
Subrata KUNDU, George Washington University

UIC Biostatistics Student Volunteers:
Cecilia CHANG, Ruizhe CHEN, Yu-Che CHUNG, Avisek DATTA, Yan GAO, Shanta GHOSH, Sasha KRAVETS, Hesen LI, Jun LU, Xiaohan MEI, Debarghya NANDI, Hai NGUYEN, Soumya SAHU, Jieqi TU, Tong ZHANG.

Sponsorships and Endorsements:
National Science Foundation, USA and Merck Sharp & Dohme, Kenilworth, NJ, USA.

Cover Design: Jyotishka DATTA, Virginia Tech.

IISA 2021

WELCOME ADDRESS

Snigdhansu Chatterjee

On behalf of the International Indian Statistical Association (IISA), we are writing to welcome you to the IISA 2021 conference. This is the official meeting of IISA and provides IISA members and conference participants a unique opportunity to meet and exchange ideas. IISA membership is open to all. The IISA is a non-profit organization, with objectives:

(i) to promote education, research and application of statistics, probability and data science throughout the world with a special emphasis on the Indian subcontinent;

(ii) to foster the exchange of information and scholarly activities between various countries as well as among other national/international organizations for the development of Statistical Science;

(iii) to serve the needs of young statisticians and data scientists;

(iv) to encourage cooperative efforts among members in education, research, industry and business.

In the tradition of past IISA conferences, IISA 2021 is providing a forum for young researchers and leading experts to discuss recent progress in statistical theory, methodology and applications and integration with data science for *Statistics in The Era of Evidence based Inference*.

The conference is hosting two vibrant student paper competitions. Two groups of judges reviewed 35 student paper submissions and 10 selected students will present their research works in two oral competition sessions. The conference is also featuring a session organized by the Caucus for Women in Statistics.

IISA 2021 is featuring three Plenary Lectures, eight Special Invited Lectures and sixty-six invited sessions with close to 220 scientific presentations. The conference speakers and registered participants encompass at least three continents! It is a huge endeavor and we acknowledge the work and efforts of the session organizers, speakers, and faculty, staff and student volunteers at the local host University of Illinois, School of Public Health. The scientific program committee, the student paper committee and other IISA committees have been working tirelessly for the conference since planning began in August 2019.

Sanjib Basu

It is a somber moment in time when all of us, but especially some parts of the world including our colleagues in India, continue to greatly suffer from the COVID pandemic. IISA 2021 is dedicating a session to our colleague Dr. Chandra who passed away in April 2021 suffering from COVID-19 and to pay respect to the sufferings from COVID pandemic in the world.

We gratefully acknowledge support of the United States National Science foundation which helped us partially support conference registrations of United States participants. We also gratefully acknowledge support from Merck Sharp & Dohme, Kenilworth, NJ, USA and local organization support from University of Illinois, School of Public Health.

We welcome you again to this exciting and stimulating IISA 2021 conference

Snigdhansu Chatterjee
President, International Indian Statistical Association
Professor, School of Statistics, University of Minnesota

Sanjib Basu
Past President (2020), International Indian Statistical Association
Paul Levy and Virginia F. Tomasek Professor of Biostatistics
School of Public Health, University of Illinois Chicago

IISA Student Paper Competition 2021

Each year IISA conducts a student paper competition in the Probability/Theory/Methodology and Applied categories. This competition is focused on the students enrolled in an MS/Ph.D. (or equivalent) program in Statistics or related fields. The winners are decided based on the submitted article where the student is the lead author and an oral presentation by the shortlisted students during the IISA conference.

For the 2021 IISA student paper competition in the area of Probability/Theory/Methodology, we received eighteen very strong submissions. The topics of the submissions encompassed varied and diverse fields such as Bayesian statistics, network analysis, probability theory, high dimensional statistics, and machine learning.

There were seventeen papers in the Applied category submitted by students from a myriad of institutions. Statistical methods were applied to various areas such as HIV/AIDS, air pollution, COVID-19, and gold futures markets. Novel study designs and analysis methods were both represented in the seventeen papers.

The judges recommended in total ten submissions, five from each category for oral presentations. These presentations will be held on May 21, 2021.

Acknowledgement:

IISA would like to thank the following judges and the chairs of committees for their time and help in conducting the student paper competition:

- Soutir BANDYOPADHYAY, Colorado School of Mines
- Samprit BANERJEE, Cornell University in NYC, New York
- Sushama BENDRE, Indian Statistical Institute, Chennai
- Sharmodeep BHATTACHARYA, Department of Statistics, Oregon State University
- Shrabanti CHOWDHURY, Mount Sinai in NYC, New York
- Sumit MUKHERJEE, Department of Statistics, Columbia University
- Bhuvanesh PAREEK, Indian Institute of Management, Indore
- Debdeep PATI, Department of Statistics, Texas A&M University
- Sujata PATIL, Cleveland Clinic
- Bodhisattva SEN, Department of Statistics, Columbia University

Contents

Programme 3

Abstracts 35

Directory 103

Instructions:

- **All times are in US Central time.**
- All sessions will be conducted live on zoom. Unfortunately, we cannot handle pre-recorded lectures.
- The Programme book contains links that can be used to navigate its pages.
- **This booklet does not contain the live zoom links. In order to access the zoom rooms please log in to the conference registrants login system using the e-mail you used for registration. The live zoom links are available on a similar booklet on the other side of the login page.**
- **You need a separate registration for the programme on Sunday**.
- We request the *speakers* and the *chair* of each *invited* and *student paper competitions* to check the zoom links for their respective sessions in advance. They should be able to check if the speakers, microphone, and the camera on their machines are working properly. Unfortunately, the screen share function won't work before the host starts the session.
- Please enter the zoom room of your session 15 minutes before your session starts and make sure that your system is working and you should be able to share your screen.

Glossary

IS	Invited Session

Programme Overview

Thursday, May 20 **3**

8:00 - 8:30 CDT
Conference Inaugaration . . . 3

8:30 - 9:30 CDT
Plenary Lecture 1 Nilanjan Chatterjee . . . 3

9:45 - 11:00 CDT
IS 1 New Advances in Estimands for the treatment effect in clinical trials . . . 3
IS 2 Recent advances in adaptive design clinical trials . . . 3
IS 3 Big Data and Modern Designs . . . 4
IS 4 Statistical methods and applications to problems related to management . . . 4
IS 5 On Some Recent Advances in Nonparametric Statistics . . . 4
IS 6 Advancements in spatio-temporal modeling under big-data framework . . . 5

11:15 - 12:30 CDT
IS 7 Big Data and Advanced Analytics in Pharmaceutical Development . . . 5
IS 8 Data Visualization . . . 5
IS 9 Advances in high-dimensional inference . . . 6
IS 10 Survey Sampling and Small Area Estimation . . . 6
IS 11 Recent Advances in High-Dimensional Time Series Analysis . . . 6
IS 12 Statistical contributions from the field of Genetics . . . 7

13:00 - 14:30 CDT
Special Invited Session 1 Siddhartha Chib and Michael Newton . . . 7
Special Invited Session 2 Mousumi Banerjee and Susan Paddock . . . 7

14:45 - 16:00 CDT
IS 13 New developments in Statistics . . . 8
IS 14 Recent Advancement in Statistical and Computational Methods for Precision Medicine . . . 8
IS 15 Bayesian Machine Learning . . . 8
IS 16 Empirical Bayes Methodology . . . 9
IS 17 Recent Advances in Estimation theory . . . 9
IS 18 Methods for Single-cell and Microbiome Sequencing Data . . . 9
IS 19 Orphan and Rare Diseases . . . 10

16:30 - 17:45 CDT

IS 20 Emerging Statistical Innovations in Clinical Research 10
IS 21 Exchangeability and its applications in Clinical trials 11
IS 22 Big data analytics with applications to astronomy 11
IS 23 Biostatistics with Bayesian Additive Regression Trees 11
IS 24 Advances in statistical methods for survival analysis 12
IS 25 Advances in Graphical Modeling 12

Friday, May 21 **13**

8:30 - 9:30 CDT

Plenary Lecture 2 Kannan Natarajan 13

9:45 - 11:00 CDT

IS 26 Drug Safety Analysis 13
IS 27 Innovative Methods in Clinical Trials 13
IS 28 Statistical inference for high dimensional data 13
IS 29 It's all about Bayes 14
IS 30 Use of Histoical Control Data and Evidence Synthesis in Clinical Trials 14
IS 31 Advances in Network Data Analysis. 14
Student Paper Competition 1 Theory and Methodology 15

11:15 - 12:30 CDT

IS 32 Some Bayesian Perspectives in Pharmaceutical Statistical Methods 15
IS 33 Design and analysis in early-stage clinical trials 16
IS 34 Advanced Statistical Methods for High-Dimensional Data . . . 16
IS 35 Bayesian Modeling and Computation 17
IS 36 Recent Developments in Network Data Inference 17
IS 37 Recent Advances in Statistical Genetics: Session Organised by the Caucus for Women in Statistics 17

13:15 - 14:30 CDT

IS 38 Precision Medicine 18
IS 39 Machine Learning Methods for High-dimensional and Functional Data 18
IS 40 Statistical and Computational Advances for Large Scale Data with High Dimensions 18
IS 41 Advances in Bayesian semiparametric and linear mixed models 19
IS 42 Variable selection and its application in biomedical sciences . . 19
Student Paper Competition 2 Applications 19

14:45 - 16:00 CDT

IS 43 Advances in Multiplicity Control Methods in Health care . . . 20
IS 44 Further exploration of the MaxCombo Test in Immuno-Oncology (IO) trials: A follow-up on the original proposal from the Non-Proportional Hazards (NPH) working group 20
IS 45 Small Area Estimation: Dedicated to the memory of Professor Hukum Chandra (1972 - 2021) 21
IS 46 Innovative Bayesian Approach to advance Drug development . 21
IS 47 Recent advances in Trend Filtering and related methods. . . . 21

IS 48 Multivariate Modeling for Improved Detection of Genetic Variants 22

16:15 - 17:30 CDT

IS 49 Multiple Statistical Approaches to Address Missing Data in Drug Development Trials 22

IS 50 Supersaturated and Sequential Designs 23

IS 51 Methods for High Dimensional Data 23

IS 52 Bayesian Methods and Applications for Complex Data 23

IS 53 Statistical Methods for Contemporary Applications 24

IS 54 Topics in Network Inference 24

Saturday, May 22 **25**

8:30 - 9:30 CDT

Plenary Lecture 3 Xihong Lin 25

9:45 - 11:15 CDT

Special Invited Session 3 Roshan Joseph and Bani Kumar Mallick 25

Special Invited Session 4 Frank Bretz and Aloka Chakravarty 25

11:30 - 12:45 CDT

IS 55 Biomarker Characterization and Development for Patient Enrichment in Oncology Clinical Trial 25

IS 56 Radomization and More 26

IS 57 Subsampling Methods in the Era of Big Data 26

IS 58 Contemporary Bayesian Modeling with Applications 26

IS 59 Advances in Random Forests and Decision Tree Ensembles 27

IS 60 Advances in Analysis of Intensive Longitudinal Data 27

13:00 - 14:15 CDT

IS 61 Novel Methods for Oncology Drug Development 27

IS 62 Efficient Designs 28

IS 63 Modeling and Analysis of High Dimensional Graphs, Networks, and Tensors 28

IS 64 Advances in High-Dimensional Bayesian Methodology 28

IS 65 Statistical methods for genetic prediction of disease using 'omics data 29

IS 66 Modern computational methods for spatial data 29

Sunday, May 23 **31**

10:00 - 10:30 CDT

In Memorium: Professor Hukum Chandra (1972 - 2021) 31

10:40 - 12:00 CDT

Meeting 1 IISA General Body Meeting 31

Programme

Thursday May 20

Conference Inaugaration *Venue:* Room A

Chair : Sanjib BASU, School of Public Health, University of Illinois Chicago
Time : **8:00 - 8:30 CDT**

- **Wayne H. GILES**, Dean, School of Public Health, University of Illinois Chicago
- **Ansu CHATTERJEE**, School of Statistics, University of Minnesota
- **Saonli BASU**, Division of Biostatistics, University of Minnesota

Plenary Lecture 1 Nilanjan Chatterjee *Venue:* Room A

Chair : Saonli BASU, Division of Biostatistics, University of Minnesota

8:30 **Individualized Risk Prediction: Lessons from Genetics and COVID-19** [**Abstract** 28]

Nilanjan CHATTERJEE, *615 N WOLFE ST, Suite E3527,Johns Hopkins University*
Jin JIN, *Johns Hopkins University*
Prosenjit KUNDU, *Johns Hopkins University*
Neha AGARWALA, *University of Maryland*
Haoyu ZHANG, *Harvard University*

IS 1 New Advances in Estimands for the treatment effect in clinical trials *Venue:* Room 1

Chair and Organizer : Shanthi SETHURAMAN, Eli Lilly and Company

9:45 **New Perspectives on Causal Inference in Clinical Trials With Multiple Endpoints, Repeated Measures and Subject Non-Adherence** [**Abstract** 186]

Hakeem WAHAB, *Statistics,Purdue University*
Stephen RUBERG, *Analytix Thinking*
Hege MICHIELS, *Ghent University*
Arman SABBAGHI, *Purdue University*

10:10 **Causal inference and estimands in clinical trials** [**Abstract** 106]

Ilya LIPKOVICH, *Real world analytics,Eli Lilly and Company*
Bohdana RATITCH, *Bayer*
Craig MALLINCKRODT, *Biogen Idec*

10:35 **Recent advances in defining estimands and imputation of missing data in clinical trials** [**Abstract** 147]

Yongming QU, *Department of Data and Analytics,Eli Lilly and Company*

IS 2 Recent advances in adaptive design clinical trials *Venue:* Room 2

Chair and Organizer : Madan KUNDU, Daiichi Sankyo Inc

9:45 **Simulation Practices for Adaptive Clinical Trial Design in Drug and Device Development** [**Abstract** 37]

Greg CICCONETTI, *Teri Anderson,AbbVie*

10:10 **A practical Response Adaptive Block Randomization (RABR) design with analytic type I error protection** [**Abstract** 204]

Tianyu ZHAN, *Data and Statistical Sciences, AbbVie Inc.,Data and Statistical Sciences, AbbVie Inc.*

10:35 **Adaptive Multi-Arm Multi-Stage Design** [Abstract 55]

Pranab GHOSH, *Biostatistics,Pfizer Inc.*

IS 3 Big Data and Modern Designs *Venue:* Room 3

Chair : Dibyen MAJUMDAR, Math., Stat. and Comp., Sci, UIC,University of Illinois at Chicago
Organizer : Abhyuday MANDAL, University of Georgia

9:45 **Collaborative Design for Improved Causal Machine Learning on Big Observational Data** [Abstract 154]

Arman SABBAGHI, *Department of Statistics,Purdue University*
Yumin ZHANG, *Purdue University Department of Statistics*

10:10 **Score-Matching Representative Approach for Big Data Analysis with Generalized Linear Models** [Abstract 198]

Jie YANG, *Department of Mathematics, Statistics, and Computer Science,University of Illinois at Chicago*
Keren LI, *Northwestern University*

10:35 **Statistical Computing Meets Quantum Computing** [Abstract 109]

Ping MA, *Statistics,University of Georgia*

IS 4 Statistical methods and applications to problems related to management *Venue:* Room 4

Chair and Organizer : Karthik SRIRAM, Indian Institute of Management Ahmedabad,India

9:45 **Distribution free estimation of optimal order quantity for a newsboy** [Abstract 131]

Sujay MUKHOTI, *Operations Management and Quantitative Techniques,Indian Institute of Management Indore*

10:10 **Bayesian time-aligned factor analysis of paired multivariate time series** [Abstract 150]

Arkaprava ROY, *University of Florida*
David DUNSON, *Duke University*
Jana SCHAICH-BORG, *Duke University*

10:35 **Reinforced designs: Multiple instruments plus control groups as evidence factors in an observational study** [Abstract 80]

Bikram KARMAKAR, *Statistics,University of Florida*
Dylan S. SMALL, *University of Pennsylvania*
Paul R. ROSENBAUM, *University of Pennsylvania*

IS 5 On Some Recent Advances in Nonparametric Statistics *Venue:* Room 5

Chair and Organizer : Bodhisattva SEN, Department of Statistics,Columbia University

9:45 **Convex Regression in High Dimensions** [Abstract 60]

Adityanand GUNTUBOYINA, *Statistics,University of California Berkeley*
Gil KUR, *Massachusetts Institute of Technology*
Fuchang GAO, *University of Idaho*
Bodhisattva SEN, *Columbia University*

10:10 **Refined Maximal Inequalities with Applications to Oracle Inequalities** [Abstract 89]

Arun KUCHIBHOTLA, *Department of Statistics and Data Science,Carnegie Mellon University*

10:35 **Least Squares Estimation of a Monotone Quasiconvex Regression Function** [Abstract 140]

Rohit PATRA, *Department of Statistics,University of Florida*
Somabha MUKHERJEE, *University of Pennsylvania*
Andrew L JOHNSON, *Amazon*
Hiroshi MORITA, *Osaka University*

IS 6 Advancements in spatio-temporal modeling under big-data framework *Venue:* Room 6

Chair and Organizer : Indranil SAHOO, Statistical Sciences & Operations Research,Virginia Commonwealth University

9:45 **Stochastic Generators with Global Spatio-Temporal Locally Diffusive SPDE Models** [Abstract 20]

Stefano CASTRUCCIO, *Stefano Castruccio,University of Notre Dame*
Geir-Arne FUGLSTAD, *Norwegian University of Science and Technology*

10:10 **Multivariate spectral downscaling for PM2.5 species** [Abstract 58]

Yawen GUAN, *Statistics,University of Nebraska - Lincoln*
Brian J REICH, *North Carolina State University*
James A MULHOLLAND, *Georgia Tech*
Howard H CHANG, *Emory University*

10:35 **Computing models with big data: privacy consideration, distributed implementation** [Abstract 59]

Rajarshi GUHANIYOGI, *Statistics, UC Santa Cruz*

IS 7 Big Data and Advanced Analytics in Pharmaceutical Development *Venue:* Room 1

Chair : Li WANG, Data and Statistical Sciences, AbbVie
Organizer : Haoda FU, Eli Lilly and Company

11:15 **Visual Analytics, Big Data, and Drug Safety with a Focus on Text Mining and Natural Language Processing** [Abstract 132]

Melvin MUNSAKA, *Statistical Sciences,AbbVie*
Kefei ZHOU, *Bristol Myers Squibb*
Krishan SINGH, *GlaxoSmithKline (Retired)*

11:40 **Analytic framework for non-randomized single-arm clinical trials with external RWD control** [Abstract 188]

Hongwei WANG, *Global Medical Affairs Statistics,AbbVie*
Yixin FANG, *AbbVie*
Weili HE, *AbbVie*

12:05 **Advanced Machine Learning to Predict Enrollment to Expedite Clinical Trials** [Abstract 196]

Yunzhao XING, *Data and Statistical Science,AbbVie*
Li WANG, *AbbVie*

IS 8 Data Visualization *Venue:* Room 2

Chair and Organizer : Vipin ARORA, Eli Lilly and Company

11:15 **Effective use of Visual Analytics in Monitoring Clinical Trial Data** [Abstract 170]

Abigail SLOAN, *Biostatistics,Pfizer*
Anindita BANERJEE, *Pfizer*

11:40 On-demand safety and efficacy insights [Abstract 72]

Cathleen JEWELL, *Clinical Analytics,AbbVie*

12:05 Techniques and methods for data visualization in clinical trials illustrated with examples [Abstract 149]

Mallikarjuna RETTIGANTI, *Neuroscience,Eli Lilly and Company*
Bochao JIA, *Eli Lilly and Company*
Eric WOLF, *Eli Lilly and Company*
Jakub JEDYNAK, *Eli Lilly and Company*

IS 9 Advances in high-dimensional inference

Venue: Room 3

Chair and Organizer : Debashis PAUL, University of California, Davis

11:15 Consistency of Spectral Clustering on Hierarchical Stochastic Block Models [Abstract 100]

Xiaodong LI, *UC Davis/Statistics,UC Davis*
Lihua LEI, *Stanford University*
Xingmei LOU, *UC Davis*

11:40 ODE backpropagation dynamics and reparameterization for high dimensional lensing inference from the CMB polarization [Abstract 2]

Ethan ANDERES, *Department of Statistics,University of California at Davis*

12:05 Improved Nonparametric Empirical Bayes Estimation using Transfer Learning [Abstract 127]

Gourab MUKHERJEE, *Department of Data Sciences & Operations,University of Southern California*

IS 10 Survey Sampling and Small Area Estimation

Venue: Room 4

Chair : Min YANG, Department of Mathematics, Statistics, and Computer Science,University of Illinois at Chicago

Organizer : Jennifer PAJDA-DE LA O, Mathematics, Statistics, and Computer Science,University of Illinois at Chicago

11:15 Median-Unbiasedness in Finite Population Survey Sampling [Abstract 137]

Jennifer PAJDA-DE LA O, *Mathematics, Statistics, and Computer Science,University of Illinois at Chicago*
A. S. HEDAYAT, *University of Illinois at Chicago*

11:40 Edge Selection for Graphical Models with Mixed Types under Informative Sampling [Abstract 12]

Emily BERG, *Statistics,Iowa State University*
Hao SUN, *Iowa State University*
Zhengyuan ZHU, *Iowa State University*

IS 11 Recent Advances in High-Dimensional Time Series Analysis

Venue: Room 5

Chair : Abhirup DATTA, Johns Hopkins University

Organizer : Sumanta BASU, Statistics and Data Science,Cornell University

11:15 Large Spectral Density Matrix Estimation by Thresholding [Abstract 11]

Sumanta BASU, *Statistics and Data Science,Cornell Uniersity*

11:40 Statistical Inference for Networks of Point Processes [Abstract **165**]

Ali SHOJAIE, *Biostatistics, University of Washington*

12:05 Multiple Change Point Detection in Reduced Rank High Dimensional Vector Autoregressive Models [Abstract **156**]

Abolfazl SAFIKHANI, *University of Florida, University of Florida*
Peiliang BAI, *University of Florida*
George MICHAILIDIS, *University of Florida*

IS 12 Statistical contributions from the field of Genetics *Venue:* Room 6

Chair and Organizer : Debashree RAY, Johns Hopkins University

11:15 Kernel Distance Covariance Approach for Testing Association in Longitudinal Studies [Abstract **153**]

Pratyaydipta RUDRA, *Statistics, Oklahoma State University*

11:40 A Log–Linear Model for Inference on Bias in Microbiome Studies [Abstract **211**]

Ni ZHAO, *Johns Hopkins University, assistant professor of Biostatistics*
Glen SATTEN, *Emory University*

12:05 Fine-tuning genetic prediction models using marginal association statistics [Abstract **108**]

Qiongshi LU, *Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison*

Break 12:30 - 13:00 CDT

Special Invited Session 1 Siddhartha Chib and Michael Newton *Venue:* Room B

Chair : Sanjay CHAUDHURI, National University of Singapore

13:00 Bayesian Estimation and Comparison of Conditional Moment Models [Abstract **33**]

Siddhartha CHIB, *Olin Business School, Washington University in Saint Louis*
Minchul SHIN, *Federal Reserve Bank, Philadelphia*
Anna SIMONI, *CREST, CNRS, Ecole Polytechnique, Paris*

13:45 Empirical Bayes and the false discovery rate, revisited [Abstract **135**]

Michael NEWTON, *Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison*

Special Invited Session 2 Mousumi Banerjee and Susan Paddock *Venue:* Room C

Chair : Pralay MUKHOPADHYAY, Department of Biometrics, Otsuka America Pharmaceuticals Inc.

13:00 Statistical Issues and Challenges in Analyzing Data from a Quality Improvement Collaborative [Abstract **6**]

Mousumi BANERJEE, *Biostatistics, University of Michigan*

13:45 Causal Inference Under Interference in Dynamic Therapy Group Studies [Abstract **136**]

Susan PADDOCK, *Statistics and Methodology, NORC at the University of Chicago*
Bing HAN, *RAND Corporation*
Lane BURGETTE, *RAND Corporation*

IS 13 New developments in Statistics

Venue: Room 1

Chair : Sanjay CHAUDHURI, National University of Singapore

Organizer : Gauri Sankar DATTA, Department of Statistics,University of Georgia/US Census Bureau

14:45 Using Household and Retail Scanner Data to Inform Food and Nutrition Policy [Abstract 213]

Chen ZHEN, *Agricultural and Applied Economics,University of Georgia*
Lan MU, *University of Georgia*
Gauri DATTA, *University of Georgia*
Chandra DHAKAL, *University of Georgia*

15:10 A Hierarchical Bayes Unit-Level Small Area Estimation Model for Normal Mixture Populations [Abstract 119]

Abhyuday MANDAL, *University of Georgia*
Gauri S DATTA, *University of Georgia*
Adrijo CHAKRABORTY, *U.S. Food and Drug Administration*
Shuchi GOYAL, *University of California, Los Angeles*

15:35 Bayesian Hierarchical Spatial Models for Small Area Estimation [Abstract 36]

Hee Cheol CHUNG, *Department of Statistics,Texas A&M University*
Gauri Sankar DATTA, *University of Georgia*

IS 14 Recent Advancement in Statistical and Computational Methods for Precision Medicine

Venue: Room 2

Chair : Yan SUN, Abbvie Inc.

Organizer : Xin HUANG, Discovery and Exploratory Statistics (DIVES), Data & Statistical Sciences,AbbVie Inc.

14:45 Examples and Stories: Computer Scientists and Statisticians - What we need to learn from each other [Abstract 51]

Haoda FU, *Eli Lilly and Company*

15:10 The two strategies in estimating the values of dynamic treatment regimes [Abstract 47]

Yixin FANG, *Data and Statistical Sciences,AbbVie*

15:35 Causal Machine Learning for predictive biomarker identification [Abstract 66]

Xin HUANG, *Discovery and Exploratory Statistics (DIVES), Data & Statistical Sciences,AbbVie Inc.*

IS 15 Bayesian Machine Learning

Venue: Room 3

Chair and Organizer : Purushottam LAUD, Medical College of Wisconsin

14:45 Causal Inference with the Instrumental Variable Approach and Bayesian Nonparametric Machine Learning [Abstract 122]

Robert MCCULLOCH, *School of Mathematical and Statistical Sciences,Arizona State*
Purushottam LAUD, *Medical College of Wisconsin*
Brent LOGAN, *Medical College of Wisconsin*
Rodney SPARAPANI, *Medical College of Wisconsin*

15:10 Convergence complexity analysis of MCMC algorithm [Abstract 146]

Qian QIN, *School of Statistics,University of Minnesota*
James P. HOBERT, *University of Florida*

15:35 Mixture cure rate models with BART [Abstract 99]

Xiao LI, *Biostatistics,Medical College of Wisconsin*
Rodney SPARAPANI, *Medical College of Wisconsin*
Brent LOGAN, *Medical College of Wisconsin*

IS 16 Empirical Bayes Methodology *Venue:* Room 4

Chair and Organizer : Adityanand GUNTUBOYINA, Statistics,University of California Berkeley

14:45 Improved Nonparametric Empirical Bayes Estimation By Transfer Learning [Abstract 177]

Wenguang SUN, *University of Southern California,Session on Empirical Bayes Methodology*
Gourab MUKHERJEE, *University of Southern California*
Jiajun LUO, *University of Southern California*

15:10 Estimating an Unknown Multi-dimensional Prior from Heterogeneous Data via Nonparametric Maximum Likelihood [Abstract 162]

Bodhisattva SEN, *Department of Statistics,Columbia University*
Jake SOLOFF, *University of California at Berkeley*
Adityanand GUNTUBOYINA, *University of California at Berkeley*

15:35 Conditional calibration for FDR control under dependence [Abstract 50]

Will FITHIAN, *Statistics, UC Berkeley,UC Berkeley Statistics*
Lihua LEI, *Stanford University*

IS 17 Recent Advances in Estimation theory *Venue:* Room 5

Chair : Bhaswar BHATTACHARYA, Statistics,University of Pennsylvania
Organizer : Gourab MUKHERJEE, Department of Data Sciences & Operations,University of Southern California

14:45 Motif Estimation via Subgraph Sampling: The Fourth-Moment Phenomenon [Abstract 14]

Bhaswar BHATTACHARYA, *Statistics,University of Pennsylvania*
Sayan DAS, *Columbia University*
Sumit MUKHERJEE, *Columbia University*

15:10 Piecewise Polynomial Estimation on Grids by Dyadic CART and Optimal Regression Tree [Abstract 29]

Sabyasachi CHATTERJEE, *Statistics,University of Illinois at Urbana-Champaign*
Subhajit GOSWAMI, *Tata Institute of Fundamental Research*

15:35 Global testing for dependent Bernoullis [Abstract 129]

Sumit MUKHERJEE, *Sumit Mukherjee,Columbia University*
Nabarun DEB, *Columbia University*
Rajarshi MUKHERJEE, *Harvard University*
Ming YUAN, *Columbia University*

IS 18 Methods for Single-cell and Microbiome Sequencing Data *Venue:* Room 6

Chair : Soutik GHOSAL, NICHD/DIPHR
Organizer : Himel MALLICK, Biostatistics and Research Decision Sciences,Merck Research Laboratories

14:45 Omics Community Detection using Multi-resolution Clustering [Abstract 148]

Ali RAHNAVARD, *Biostatistics and Bioinformatics,George Washington University*
Himel MALLICK, *Merck & Co., Inc.*
Suvo CHATTERJEE, *National Institutes of Health*
Bahar SAYOLDIN, *George Mason University*
Keith A. CRANDALL, *George Washington University*

15:10 Differential expression of single-cell RNA-seq data using Tweedie models [Abstract 117]

Himel MALLICK, *Biostatistics and Research Decision Sciences,Merck Research Laboratories*
Suvo CHATTERJEE, *National Institutes of Health*
Shrabanti CHOWDHURY, *Icahn School of Medicine at Mount Sinai*
Sapatarshi CHATTERJEE, *Eli Lilly & Company*
Ali Rahnavard, Stephanie HICKS, *George Washington University, Johns Hopkins University*

15:35 Learning directed acyclic graphs based on single cell RNAseq data [Abstract 35]

SHRABANTI CHOWDHURY, *Genetics and Genomic Sciences,Icahn school of Medicine at Mount Sinai*

IS 19 Orphan and Rare Diseases *Venue:* Room 7

Chair and Organizer : Hrishikesh KULKARNI, Alexion Pharmaceuticals

14:45 Statistical Considerations in Clinical Trial Design for Rare Diseases [Abstract 193]

Zailong WANG, *AbbVie,AbbVie*
Lanju ZHANG, *AbbVie*

15:10 The Role of Natural History Data in Rare Disease Clinical Development [Abstract 138]

Jeff PALMER, *Early Clinical Development Statistics,Pfizer, Inc.*

15:35 Efficient Trial Design Selection in Rare Disease Clinical Development [Abstract 71]

Yannis JEMIAI, *Cytel,Cytel*

IS 20 Emerging Statistical Innovations in Clinical Research *Venue:* Room 1

Chair : Li WANG, Data and Statistical Science, AbbVie
Organizer : Saurabh MUKHOPADHYAY, Data and Statistical Science, AbbVie

16:30 Digital endpoint development in atopic dermatitis clinical trials: using deep learning to quantify nocturnal scratch based on actigraphy data [Abstract 31]

Xiaotian CHEN, *Data and Statistical Science, AbbVie*
Li WANG, *Data and Statistical Science, AbbVie*

16:55 Bayesian Interim Decision Strategy in Phase III Time-To-Event Study [Abstract 68]

David IPE, *Data and Statistical Science, AbbVie*
Saurabh MUKHOPADHYAY, *Data and Statistical Science, AbbVie*

17:20 Discussant: Saurabh MUKHOPADHYAY, Data and Statistical Science, AbbVie

IS 21 Exchangeability and its applications in Clinical trials *Venue:* Room 2

Chair and Organizer : Tanujit DEY, Center for Surgery and Public Health, BWH,Harvard Medical School

16:30 Bayesian Basket Trial Design with False Discovery Rate Control [Abstract 203]

Emily ZABOR, *Department of Quantitative Health Sciences, Cleveland Clinic*
Brian HOBBS, *The University of Texas at Austin*
Michael KANE, *Yale University*
Satrajit ROYCHOUDHURY, *Pfizer, Inc*
Lei NI, *U.S. Food and Drug Administration*

16:55 Bayesian Design of Pediatric Clinical Trials with Prospective Incorporation of Data from Adult Trials [Abstract 144]

Matthew PSIODA, *Department of Biostatistics,Department of Biostatistics*

17:20 Sequential basket trial design based on multi-source exchangeability with predictive probability monitoring [Abstract 77]

Alexander KAIZER, *Biostatistics and Informatics,University of Colorado-Anschutz Medical Campus*
Emily ZABOR, *Cleveland Clinic*
Nan CHEN, *Gilead Sciences*
Brian HOBBS, *University of Texas-Austin*

IS 22 Big data analytics with applications to astronomy *Venue:* Room 3

Chair and Organizer : Samiran SINHA, Texas A&M University

16:30 Adaptive Methods for Time-Modulated Variable Stars [Abstract 125]

Giovanni MOTTA, *Statistics,Texas A&M, Department of Statistics*

16:55 X-ray Astronomy in 4 Dimensions [Abstract 82]

Vinay KASHYAP, *High-Energy Astrophysics Division, Center for Astrophysics – Harvard & Smithsonian*
CHASC CHASC, *CHASC*

17:20 Improving Exoplanet Detection Power: Multivariate Gaussian Process Models for Stellar Activity [Abstract 74]

David JONES, *Department of Statistics,Texas A&M University*
David C. STENNING, *Simon Fraser University*
Eric B. FORD, *Penn State University*
Robert L WOLPERT, *Duke University*
Thomas J LOREDO, *Cornell University*

IS 23 Biostatistics with Bayesian Additive Regression Trees *Venue:* Room 4

Chair : Purushottam LAUD, Medical College of Wisconsin
Organizer : Rodney SPARAPANI, Division of Biostatistics,Medical College of Wisconsin

16:30 TransPRECISE: Personalized Network Modeling of the Pan-cancer Patient and Cell Line Interactome [Abstract 16]

Rupam BHATTACHARYYA, *Biostatistics,University of Michigan*
Min Jin HA, *The University of Texas MD Anderson Cancer Center*
Qingzhi LIU, *University of Michigan*
Rehan AKBANI, *The University of Texas MD Anderson Cancer Center*
Han LIANG, *The University of Texas MD Anderson Cancer Center*

16:55 **Random Effects with Bayesian Additive Regression Trees for Precision Medicine** [**Abstract** 171]

Charles SPANBAUER, *Biostatistics,University of Minnesota*
Rodney SPARAPANI, *Medical College of Wisconsin*

17:20 **Nonparametric Failure Time with BART and DPM LIO** [**Abstract** 172]

Rodney SPARAPANI, *Division of Biostatistics,Medical College of Wisconsin*
Laud, PURUSHOTTAM, *Medical College of Wisconsin*
Logan, BRENT, *Medical College of Wisconsin*
McCulloch, ROBERT, *Arizona State University*
Pratola, MATT, *Ohio State University*

IS 24 Advances in statistical methods for survival analysis *Venue:* Room 5

Chair and Organizer : Sunyoung SHIN, University of Texas at Dallas

16:30 **Efficient competing risks regression models under the generalized case-cohort design** [**Abstract** 84]

Soyoung KIM, *Division of Biostatistics,Medical College of Wisconsin*
Yayun XU, *Merck Pharmaceutical company,*
Mei-Jie Zhang, Kwang Woo AHN, *Division of Biostatistics, Medical College of Wisconsin*
David COUPER, *Department of Biostatistics, University of North Carolina at Chapel Hill*

16:55 **On recurrent-event win ratio** [**Abstract** 120]

Lu MAO, *Biostatistics and Medical Informatics,University of Wisconsin-Madison*
KyungMann KIM, *University of Wisconsin-Madison*
Yi LI, *University of Wisconsin-Madison*

17:20 **Survival Trees for Informative Interval-Censored Data** [**Abstract** 67]

Noorie HYUN, *Division of Biostatistics,Medical College of Wisconsin*
Xiao LI, *Medical College of Wisconsin*

IS 25 Advances in Graphical Modeling *Venue:* Room 6

Chair and Organizer : Lin ZHANG, Division of Biostatistics, University of Minnesota

16:30 **Bayesian Integrative Approaches to Enable Precision Medicine** [**Abstract** 4]

Veera BALADANDAYUTHAPANI, *Biostatistics,University of Michigan*
Min Jin Ha, Abhisek Saha, Satwik ACHARYYA, *MD Anderson, NIH, University of Michigan*

16:55 **Bayesian Spatial Blind Source Separation via the Thresholded Gaussian Process** [**Abstract** 78]

Jian KANG, *Jian Kang,University of Michigan*
Ben WU, *Renmin University*
Ying GUO, *Emory University*

17:20 **Bi-level graphical modeling of functional connectivity analysis of resting-state fMRI data** [**Abstract** 207]

Lin ZHANG, *Division of Biostatistics, University of Minnesota*

Friday May 21

Plenary Lecture 2 Kannan Natarajan

Venue: Room D

Chair : Amarjot KAUR, Merck & Co., Inc.

8:30 **Drug Development in the 21st Century Need for Innovation in Statistical Thinking** [**Abstract** 134]

Kannan NATARAJAN, *Global Head of Biometrics and Data Management,Pfizer*

IS 26 Drug Safety Analysis

Venue: Room 1

Chair and Organizer : Hrishikesh KULKARNI, Alexion Pharmaceuticals

9:45 **Aggregate Safety Analysis and Planning in Clinical Development** [**Abstract** 63]

Barbara HENDRICKSON, *Pharmacovigilance and Patient Safety, AbbVie,AbbVie*

10:10 **TITE-BOIN-ET: Time-to-event Bayesian optimal interval design to accelerate dose-finding based on both efficacy and toxicity outcomes** [**Abstract** 180]

Kentaro TAKEDA, *Data Science,Astellas Pharma Global Development, Inc*
Satoshi MORITA, *Kyoto University*
Masataka TAGURI, *Yokohama City University*

10:35 **The Full Picture: Making the Data Insight a Reality** [**Abstract** 65]

Erya HUANG, *Clinical Statistics,Bayer U.S. LLC*

IS 27 Innovative Methods in Clinical Trials

Venue: Room 2

Chair and Organizer : Rianka BHATTACHARYA, Abbvie Inc.

9:45 **REVIEW ON MODEL DIAGNOSTICS FOR INCOMPLETE DATA** [**Abstract** 27]

Arkendu CHATTERJEE, *Global Biometrics and Data Science, BMS,Associate Director, BMS*

10:10

New Approaches for Testing Non-inferiority for Three-arm Trials with Poisson Distributed Outcomes [**Abstract** 141]

Erina PAUL, *Biostatistics and Research Decision Sciences,Merck & Co., Inc.*
Samiran GHOSH,
Shrabanti CHOWDHURY,
Ram TIWARI,

10:35 **Assessing contribution of treatment phases through tipping point analyses using rank preserving structural failure time models** [**Abstract** 15]

Sudipta BHATTACHARYA, *SQS, Biostatistics,Takeda*
Jyotirmoy DEY, *Regeneron*

IS 28 Statistical inference for high dimensional data

Venue: Room 3

Chair and Organizer : Ping-Shou ZHONG, University of Illinois at Chicago

9:45 **Inference for Differential Networks in a High-dimensional Setting** [**Abstract** 86]

Mladen KOLAR, *The University of Chicago Booth School of Business,The University of Chicago Booth School of Business*
Byol KIM, *Irina Gaynanova*

10:10 **Causal inference in Mendelian Randomization with weak and heterogenous instruments [Abstract 189]**

Jingshu WANG, *Statistics,University of Chicago*
Qingyuan ZHAO, *University of Cambridge*
Jack BOWDEN, *The University of Exeter*
Dylan SMALL, *University of Pennsylvania*
Nancy R. ZHANG, *University of Pennsylvania*

10:35 **High-dimensional quadratic classifiers under the strongly spiked eigenvalue model [Abstract 69]**

Aki ISHII, *Department of Information Sciences,Tokyo University of Science*
Kazuyoshi YATA, *Institute of Mathematics, University of Tsukuba*
Makoto AOSHIMA, *Institute of Mathematics, University of Tsukuba*

IS 29 It's all about Bayes *Venue:* Room 4

Chair : Sanjib BASU, School of Public Health, University of Illinois Chicago
Organizer : Ananda SEN, Department of Biostatistics, University of Michigan

9:45 **Bayesian Model Assessment and Selection Using Bregman Divergence [Abstract 44]**

Dipak DEY, *Department of Statistics,University of Connecticut*
Gyuhyeong GOH, *Kansas State University*

10:10 **A Bayesian Joint Model for Clustered Agreement Data [Abstract 161]**

Ananda SEN, *Department of Biostatistics, University of Michigan*
Wen YE, *University of Michigan*
Pin LI, *Henry Ford Health System*

IS 30 Use of Histoical Control Data and Evidence Synthesis in Clinical Trials *Venue:* Room 5

Chair and Organizer : Pabak MUKHOPADHYAY, Daiichi-Sankyo Inc

9:45 **Considerations in using External Control from Real-world Data to Support FDA Approvals [Abstract 48]**

Dai FENG, *GMA Statistics, DSS, AbbVie,AbbVie*
Meijing WU, *AbbVie*
Hongwei WANG, *AbbVie*
Yixin FANG, *AbbVie*
Weili HE, *AbbVie*

10:10 **Building an external control arm for development of a new molecular entity [Abstract 112]**

Antara MAJUMDAR, *Statistical Innovations Group,Medidata Acorn AI*
Ruthie DAVI, *Medidata Acorn AI*

10:35 **Use of External Controls in Oncology Trials [Abstract 151]**

Pourab ROY, *Office of Biostatistics,FDA*

IS 31 Advances in Network Data Analysis. *Venue:* Room 6

Chair : Debashis MONDAL, Oregon State University
Organizer : Shirshendu CHATTERJEE, City University of New York

9:45 **Optimal offline changepoint estimation in network sequences** [**Abstract** 128]

Soumendu Sundar MUKHERJEE, *Interdisciplinary Statistical Research Unit (ISRU),Indian Statistical Institute, Kolkata*
Sharmodeep BHATTACHARYYA, *Oregon State University*
Shirshendu CHATTERJEE, *City University of New York*
Trisha DAWN, *Indian Statistical Institute, Kolkata*
Tamojit SADHUKHAN, *Indian Statistical Institute, Kolkata*

10:10 **Trading off Accuracy for Speedup: Multiplier Bootstraps for Subgraph Counts** [**Abstract** 159]

Purnamrita SARKAR, *Department of Statistics and Data Sciences,Asst. Prof.*
Qiaohui LIN, *Student, UT Austin*
Robert LUNDE, *Postdoctoral Scholar, UT Austin*

10:35 **Adjusted chi-square test for degree-corrected block models** [**Abstract** 1]

Arash AMINI, *Statistics,UCLA*
Linfan ZHANG, *UCLA*

Student Paper Competition 1 Theory and Methodology

Venue: Room 7

Chair : Bodhisattva SEN, Department of Statistics,Columbia University

9:45 **Wilson loop expectations for finite gauge groups** [**Abstract** 19]

Sky CAO, *Stanford Statistics,Stanfod University*

10:00 **A Bayesian framework for sparse estimation in high-dimensional mixed frequency Vector Autoregressive models** [**Abstract** 21]

Nilanjana CHAKRABORTY, *Statistics,University of Florida*
Kshitij KHARE, *Department of Statistics, University of Florida*
George MICHAILIDIS, *Department of Statistics and the Informatics Institute, University of Florida*

10:15 **Motif Estimation via Subgraph Sampling** [**Abstract** 40]

Sayan DAS, *Department of Mathematics,Columbia University*
Bhaswar B. BHATTACHARYA, *University of Pennsylvania*
Sumit MUKHERJEE, *Columbia University*

10:30 **Confident predictions even when distributions shift** [**Abstract** 62]

Suyash GUPTA, *Statistics,Ph.D. student, Statistics, Stanford University*
Maxime CAUCHOIS, *Ph.D. student, Statistics, Stanford University*
Alnur ALI, *Post doctoral candidate, Electrical Engineering, Stanford University*
John DUCHI, *Assistant Professor, Statistics and Electrical Engineering, Stanford University*

10:45 **PROFIT: Projection-based Test in Longitudinal Functional Data** [**Abstract** 87]

Koner SALIL, *North Carolina State University*
So Young PARK,
Ana-Maria STAICU,

IS 32 Some Bayesian Perspectives in Pharmaceutical Statistical Methods

Venue: Room 1

Chair and Organizer : Arnab MAITY, Pfizer

11:15 **Practical considerations of using historical data in clinical trials: when, what and how much do we borrow?** [**Abstract** 191]

Ling WANG, *Worldwide Research, Development and Medical,Pfizer*

11:40 Statistical Construct of Extrapolation: Composite Likelihood and Bayesian Approaches [Abstract 52]

Margaret GAMALO, *GBDM-Inflammation and Immunology,Pfizer*

12:05 Bayesian Emax dose response modeling [Abstract 182]

Neal THOMAS, *Pfizer, Groton CT USA,Pfizer*

IS 33 Design and analysis in early-stage clinical trials *Venue:* Room 2

Chair and Organizer : Min YANG, Department of Mathematics, Statistics, and Computer Science,University of Illinois at Chicago

11:15 BOIN12: Bayesian Optimal Interval Phase I/II Trial Design for Utility-Based Dose Finding in Immunotherapy and Targeted Therapies [Abstract 202]

Ying YUAN, *Biostatistics,University of Texas MD Anderson Cancer Center*
Ruitao LIN, *University of Texas MD Anderson Cancer Center*
Yanhong ZHOU, *University of Texas MD Anderson Cancer Center*
Fangrong YAN, *China Pharmaceutical University*
Daniel LI, *Bristol-Myers Squibb*

11:40 Optimal Design Theory in Early-Phase Dose-Finding Problems with Late-onset Toxicity [Abstract 183]

Tian TIAN, *Biostatistics,BeiGene*
Min YANG, *University of Illinois at Chicago*

12:05 Proof of concept and dose estimation with binary responses [Abstract 199]

Min YANG, *Department of Mathematics, Statistics, and Computer Science,University of Illinois at Chicago*

IS 34 Advanced Statistical Methods for High-Dimensional Data *Venue:* Room 3

Chair : Abhijit MANDAL, Department of Mathematical Sciences, University of Texas at El Paso
Organizer : Anirban MONDAL, Department of Mathematics, Applied Mathematics, and Statistics,Case Western Reserve University

11:15 On High Dimensional, Robust, Unsupervised Record Linkage [Abstract 26]

Ansu CHATTERJEE, *School of Statistics, University of Minnesota*

11:40 Computer model emulation for high dimensional functional output from satellite remote sensing [Abstract 124]

Anirban MONDAL, *Department of Mathematics, Applied Mathematics, and Statistics,Case Western Reserve University*
Pulong MA, *SAMSI, Duke University*
Jonathan HOBBS, *Jet Propulsion Laboratory*
Emily KANG, *University of Cincinnati*
Konomi, BLEDAR, *University of Cincinnati*

12:05 Robust Variable Selection Criteria for the Penalized Regression [Abstract 118]

Abhijit MANDAL, *Department of Mathematical Sciences, University of Texas at El Paso*
Samiran GHOSH, *Wayne State University*

IS 35 Bayesian Modeling and Computation *Venue:* Room 4

Chair and Organizer : Joyee GHOSH, Statistics and Actuarial Science,The University of Iowa

11:15 Bayesian profiling multiple imputation for missing hemoglobin values in electronic health records [Abstract 166]

Yajuan SI, *Survey Research Center,University of Michigan*
Mari PALTA, *University of Wisconsin-Madison*
Maureen SMITH, *University of Wisconsin-Madison*

11:40 MCMC algorithms for Bayesian generalized linear mixed models [Abstract 152]

Vivekananda ROY, *Department of Statistics,Iowa State University*

12:05 Bayesian Modeling of North Atlantic Tropical Cyclone Activity [Abstract 54]

Joyee GHOSH, *Statistics and Actuarial Science,The University of Iowa*
Xun LI, *The University of Iowa*
Gabriele VILLARINI, *The University of Iowa*

IS 36 Recent Developments in Network Data Inference *Venue:* Room 5

Chair and Organizer : Srijan SENGUPTA, Statistics,North Carolina State University

11:15 Mixed membership stochastic blockmodels for heterogeneous networks [Abstract 32]

Yuguo CHEN, *Department of Statistics,University of Illinois at Urbana-Champaign*

11:40 Modeling continuous-time networks of relational events [Abstract 142]

Subhadeep PAUL, *Department of Statistics,The Ohio State University*
Kevin XU, *University of Toledo*
Makan ARASTUIE, *University of Toledo*

12:05 A nonparametric test of co-spectrality in networks [Abstract 163]

Srijan SENGUPTA, *Statistics,North Carolina State University*

IS 37 Recent Advances in Statistical Genetics: Session Organised by the Caucus for Women in Statistics *Venue:* Room 6

Chair and Organizer : Swati BISWAS, Mathematical Sciences,University of Texas at Dallas

11:15 SBL: Bayesian Lasso for detecting haplotypes associated with survival traits [Abstract 104]

Shili LIN, *Statistics,Ohio State University*

11:40 Correcting population stratification in association studies of rare genetic variants using generalized PCA [Abstract 184]

Asuman TURKMEN, *Department of Statistics,OHIO STATE UNIVERSITY*
Nedret BILLOR, *Auburn University*
Yuan YUAN, *Auburn University*

12:05 Efficient SNP-based Heritability Estimation using Gaussian Predictive Process in Large-scale Cohort Studies [Abstract 10]

Saonli BASU, *Division of Biostatistics, University of Minnesota*
Souvik SEAL, *Colorado School of Public Health*
Abhirup DATTA, *Johns Hopkins University*

Break 12:30 - 13:15 CDT

IS 38 Precision Medicine *Venue:* Room 1

Chair and Organizer : Rong LIU, Bristol Myers Squibb

13:15 Comparative Biomarker Modeling for Optimal Patient Selection in Immuno-oncology [Abstract 95]

Jae LEE, *Biomarker Group, Global Biometric and Data Sciences,Bristol Myers Squibb*

13:40 Innovative Statistical Thinking in Support of Translational Science and Companion Diagnostic Development [Abstract 81]

Maha KARNOUB, *Maha Karnoub,Daiichi Sankyo - Translational Medicine Biostatistics*

14:05 Adaptation Development Strategy for Oncology Biomarker-Driven Registration Trial [Abstract 212]

Xin ZHAO, *Oncolgoy Statistics,Janssen Pharceuticals*
Sudhakar RAO, *Janssen Pharceuticals*

IS 39 Machine Learning Methods for High-dimensional and Functional Data *Venue:* Room 2

Chair and Organizer : Swarnali BANERJEE, Department of Mathematics and Statistics, Loyola University Chicago

13:15 High-dimensional rank-based inference for testing relative effects [Abstract 88]

Xiaoli KONG, *Department of Mathematics and Statistics,Loyola University Chicago*
Solomon HARRAR, *University of Kentucky*

13:40 Completion and Classification of Partially Observed Curves with Application to Classification of Bovid Teeth [Abstract 121]

Gregory MATTHEWS, *Mathematics and Statistics,Loyola University Chicago*
Ofer HAREL,
Karthik BHARATH,
Sebastian KURTEK,
Juliet BROPHY,

14:05 Minimum Cost-Compression Risk in Principal Component Analysis [Abstract 7]

Swarnali BANERJEE, *Department of Mathematics and Statistics, Loyola University Chicago*
Bhargab CHATTOPADHYAY, *Indian Institute of Management Visakhapatnam*

IS 40 Statistical and Computational Advances for Large Scale Data with High Dimensions

Venue: Room 3
Chair : Yuan ZHANG, Statistics,The Ohio State University
Organizer : Naveen Naidu NARISETTY, Statistics,University of Illinois at Urbana-Champaign

13:15 Fast and accurate computation of large-scale kernel ridge regression [Abstract 200]

Yun YANG, *Statistics,University of Illinois Urbana-Champaign*

13:40 Bootstrapping Lp-Statistics in High Dimensions [Abstract 57]

Alexander GIESSING, *Department of Operations Research and Financial Engineering,Princeton University*
Jianqing FAN, *Princeton University*

14:05 Bayesian Multiple Quantile Regression Using a Score Likelihood [Abstract 133]

Naveen Naidu NARISETTY, *Statistics,University of Illinois at Urbana-Champaign*
Teng WU, *University of Illinois at Urbana-Champaign*

IS 41 Advances in Bayesian semiparametric and linear mixed models *Venue:* Room 4

Chair and Organizer : Anirban MONDAL, Department of Mathematics, Applied Mathematics, and Statistics,Case Western Reserve University

13:15 Simultaneous Selection of Multiple Important Single Nucleotide Polymorphisms in Familial Genome Wide Association Studies Data [Abstract 114]

Subhabrata MAJUMDAR, *Data science and AI Research,University of Minnesota*
Saonli BASU, *University of Minnesota*
Matt MCGUE, *University of Minnesota*
Snigdhansu CHATTERJEE, *University of Minnesota*

13:40 Bayesian Semiparametric Longitudinal Functional Mixed Models with Locally Informative Predictors [Abstract 158]

Abhra SARKAR, *Statistics and Data Sciences,The University of Texas at Austin*
Giorgio PAULON, *The University of Texas at Austin*
Peter MUELLER, *The University of Texas at Austin*

14:05 Modeling Heterogeneity in Consumer Preferences using Bayesian Methods [Abstract 92]

Choudur LAKSHMINARAYAN, *Teradata Labs,Teradata Labs*

IS 42 Variable selection and its application in biomedical sciences *Venue:* Room 5

Chair : Emily ZABOR, Department of Quantitative Health Sciences, Cleveland Clinic
Organizer : Tanujit DEY, Center for Surgery and Public Health, BWH,Harvard Medical School

13:15 Semi-parametric Bayes Regression and Variable Selection Using Network Valued Covariates. [Abstract 91]

Suprateek KUNDU, *Biostatistics,Emory University*
Xin MA, *Emory University*
Jennifer STEVENS, *Emory University*

13:40 SCALABLE BAYESIAN VARIABLE SELECTION AND GROUPINGFOR BINARY AND MULTICLASS OUTCOME DATA [Abstract 22]

Sounak CHAKRABORTY, *Statistics,University of Missouri-Columbia*

14:05 EM based approach for analysis of multi-platform genomics data [Abstract 45]

Tanujit DEY, *Center for Surgery and Public Health, BWH,Harvard Medical School*
Sounak CHAKRABORTY, *Department of Statistics, University of Missouri*
Hao XUE, *Department of Biostatistics, Harvard School of Public Health*

Student Paper Competition 2 Applications *Venue:* Room 6

Chair : Sujata PATIL, Cleveland Clinic

13:15 Adaptive and powerful microbiome multivariate association analysis via feature selection [Abstract 5]

Kalins BANERJEE, *Department of Public Health Sciences,Pennsylvania State University*
Jun CHEN, *Division of Biomedical Statistics and Informatics, Mayo Clinic*
Xiang ZHAN, *Department of Public Health Sciences, Pennsylvania State University*

13:30 Connectivity Regression [Abstract 43]

Neel M. DESAI, *Rice University*
Veera BALADANDAYUTHAPANI,
Jeffrey MORRIS,

13:45 mbImpute: an accurate and robust imputation method for microbiome data [Abstract 73]

Ruochen JIANG, *Statistics,University of California, Los Angeles*
Vivian Wei LI, *Rutgers School of Public Health*
Jessica Jingyi LI, *University of California, Los Angeles*

14:00 An empirical Bayes approach to estimating dynamic models of co-regulated gene expression [Abstract 185]

Sara VENKATRAMAN, *Department of Statistics and Data Science,Cornell University, Department of Statistics and Data Science*
Sumanta BASU, *Cornell University, Department of Statistics and Data Science*
Myung Hee LEE, *Weill Cornell Medical College, Department of Medicine*
Martin T. WELLS, *Cornell University, Department of Statistics and Data Science*

14:15 An empirical Bayes approach to estimating dynamic models of co-regulated gene expression [Abstract 192]

Selena WANG, *Ohio State University*
Subhadeep PAUL, *Ohio State University*

IS 43 Advances in Multiplicity Control Methods in Health care *Venue:* Room 1

Chair : Satya Ravi SIDDANI, Abbvie Inc.
Organizer : Satrajit ROYCHOUDHURY, Pfizer Inc.

14:45 A Modified Graphical Approach with Generalized Sequentially Rejective Principle to Control Familywise Error Rate [Abstract 61]

Wenge GUO, *Mathematical Sciences,New Jersey Institute of Technology*
Li YU, *Merck & Co.*

15:10 Group Sequential Holm and Hochberg Procedures [Abstract 181]

Ajit TAMHANE, *Ajit Tamhane,Northwestern University*
Jiangtao GOU, *Villanova University*
Dong XI, *Novartis Pharmaceuticals*

15:35 Discussant: Pabak MUKHOPADHYAY, Daiichi-Sankyo Inc

IS 44 Further exploration of the MaxCombo Test in Immuno-Oncology (IO) trials: A follow-up on the original proposal from the Non-Proportional Hazards (NPH) working group *Venue:* Room 2

Chair and Organizer : Pralay MUKHOPADHYAY, Department of Biometrics,Otsuka America Pharmaceuticals Inc.

14:45 Do MaxCombo and Weighted Logrank Tests Control Type I Error? [Abstract 3]

Keaven ANDERSON, *Methodology Research,Merck & Co., Inc.*
Pralay MUKHOPADHYAY, *Department of Biometrics,Otsuka America Pharmaceuticals Inc.*
Satrajit ROYCHOUDHURY, *Pfizer inc.*

15:10 Possible Hazards of Proportional Hazards Models [Abstract 56]

Sujit GHOSH, *Department of Statistics,North Carolina State University*
Alvin SHENG, *North Carolina State University*

15:35 **Evaluation of Logrank and MaxCombo test in Immuno-Oncology Trials A Retrospective Analysis in Patients Treated with Anti-PD1/PD-L1 Agents across Solid Tumors [Abstract 130]**

Pralay MUKHOPADHYAY, *Department of Biometrics,Otsuka America Pharmaceuticals Inc.*
Jiabu YE, *AstraZeneca*

IS 45 Small Area Estimation: Dedicated to the memory of Professor Hukum Chandra (1972 - 2021) *Venue:* Room 3

Chair : Sanjay CHAUDHURI, National University of Singapore
Organizer : Gauri Sankar DATTA, Department of Statistics,University of Georgia/US Census Bureau

14:45 **Bayesian Analysis of the Covariance Matrix of a Large Dimensional Multivariate Normal Distribution with Shrinkage Inverse Wishart Priors [Abstract 175]**

Dongchu SUN, *Statistics,University of Nebraska-Lincoln*
James O. BERGER, *Duke University*
Chengyuan SONG, *East China NOrmal University*

15:10 **An Accurate Coreset Methodology for Efficient Reduction of Spatial Data [Abstract 42]**

Ranadeep DAW, *University of Missouri, Department of Statistics,University of Missouri*
Christopher K WIKLE, *University of Missouri*

15:35 **Pseudo-Bayes Small Area Estimation via Compromise Regression Weights [Abstract 41]**

Gauri Sankar DATTA, *Department of Statistics,University of Georgia/US Census Bureau*
Lee JUHYUNG, *University of Georgia*
Li JIACHENG, *University of Georgia*

IS 46 Innovative Bayesian Approach to advance Drug development *Venue:* Room 4

Chair : Alex LIU, Bayer
Organizer : Rong LIU, Bristol Myers Squibb

14:45 **BOIN12: Bayesian Optimal Interval Phase I/II Trial Design for Utility-Based Dose Finding in Immunotherapy and Targeted Therapies [Abstract 103]**

Ruitao LIN, *Ruitao Lin,The University of Texas MD Anderson Cancer Center*

15:10 **On the Implementation of Robust Meta-Analytical-Predictive Prior [Abstract 206]**

Hongtao ZHANG, *Global Biometrics and Data Sciences,Bristol Myers Squibb*
Alan Y CHIANG, *Bristol Myers Squibb*
Mike BRANSON, *UCB Pharma*

15:35 **A framework of Bayesian optimal phase II (BOP2) clinical trial design [Abstract 215]**

Heng ZHOU, *Biostatistics and Research Decision Sciences,Merck & Co., Inc*
Ying YUAN, *MD Anderson Cancer Center*
Linda SUN, *Merck & Co., Inc*
Cong CHEN, *Merck & Co., Inc*

IS 47 Recent advances in Trend Filtering and related methods. *Venue:* Room 5

Chair and Organizer : Sabyasachi CHATTERJEE, Statistics,University of Illinois at Urbana-Champaign

14:45 Element-wise estimation error of a total variation regularized estimator for change point detection [Abstract 208]

Teng ZHANG, *Teng Zhang,University of Central Florida*

15:05 Public health data and trend filtering [Abstract 164]

James SHARPNACK, *Statistics Department,UC Davis*

15:25 Quantile trend filtering [Abstract 111]

OSCAR HERNAN MADRID PADILLA, *Statistics,University of California, Los Angeles*

15:45 Trend filtering with sub-exponential noise and for exponential families [Abstract 155]

Veeranjaneyulu SADHANALA, *Booth School of Business,University of Chicago*
Robert BASSETT, *Naval Postgraudate School*
James SHARPNACK, *University of California, Davis*
Dan MCDONALD, *University of British Columbia, Vancouver*

IS 48 Multivariate Modeling for Improved Detection of Genetic Variants *Venue:* Room 6

Chair and Organizer : Saonli BASU, Division of Biostatistics, University of Minnesota

14:45 Integrating GWAS and multi-omics QTL summary statistics to elucidate disease genetic mechanisms via a hierarchical low-rank model [Abstract 30]

Lin CHEN, *Department of Public Health Sciences,University of Chicago*
Yihao LU, *University of Chicago*
Jin LIU, *University of Chicago*

15:10 Detecting rare haplotype association with two correlated phenotypes of binary and continuous types [Abstract 17]

Swati BISWAS, *Mathematical Sciences,University of Texas at Dallas*
Xiaochen YUAN, *University of Texas at Dallas*

15:35 Bayesian network models for integrating genetics and metabolomics data [Abstract 160]

Denise SCHOLTENS, *Department of Preventive Medicine - Biostatistics,Department of Preventive Medicine - Biostatistics*

IS 49 Multiple Statistical Approaches to Address Missing Data in Drug Development Trials *Venue:* Room 1

Chair and Organizer : Arnab MAITY, Pfizer

16:15 Fitting Proportional Odds Model with Missing Responses When the Missing Data Are Nonignorable [Abstract 143]

Vivek PRADHAN, *Statistics, ECD I&I,Pfizer Inc*

16:40 Missing data: sensitivity analysis or supplementary analysis? [Abstract 176]

Steven SUN, *Statistical Decision Science,Janssen R&D*

17:05 Composite responder rate estimation under non-ignorable missingness [Abstract 90]

Madan KUNDU, *Daiichi Sankyo Inc*

IS 50 Supersaturated and Sequential Designs

Venue: Room 2

Chair : Gauri Sankar DATTA, Department of Statistics,University of Georgia/US Census Bureau
Organizer : Abhyuday MANDAL, University of Georgia

16:15 Group Orthogonal Supersaturated Designs [Abstract 113]

Dibyen MAJUMDAR, *Math., Stat. and Comp., Sci, UIC,University of Illinois at Chicago*

16:40 Musings about supersaturated designs [Abstract 174]

John STUFKEN, *Informatics and Analytics,University of North Carolina at Greensboro*
Rakhi SINGH, *University of North Carolina at Greensboro*

17:05 Optimal Product Design by Sequential Experiments in High Dimensions [Abstract 75]

Mingyu (Max) JOO, *Marketing,UC Riverside*
Thompson MICHAEL,
Allenby GREG M., *Ohio State University*

IS 51 Methods for High Dimensional Data

Venue: Room 3

Chair and Organizer : Joyee GHOSH, Statistics and Actuarial Science,The University of Iowa

16:15 Distributed Bayesian Varying Coefficient Modeling Using a Gaussian Process Prior [Abstract 173]

Sanvesh SRIVASTAVA, *Department of Statistics and Actuarial Science,The University of Iowa*
Rajarshi GUHANIYOGI, *UC Santa Cruz*
Cheng LI, *National University of Singapore*
Terrance SAVITSKY, *Bureau of Labor Statistics*

16:40 Novel dynamic multiscale spatiotemporal models for multivariate Gaussian data with applications to stratospheric temperatures [Abstract 49]

Marco FERREIRA, *Marco Ferreira,Department of Statistics, Virginia Tech*
Mohamed ELKHOULY, *Department of Statistics, University of Wisconsin - Madison*

17:05 Variable selection in mixture of regression models: Uncovering cluster structure and relevant features [Abstract 179]

Mahlet TADESSE, *Department of Mathematics and Statistics,Georgetown University*

IS 52 Bayesian Methods and Applications for Complex Data

Venue: Room 4

Chair and Organizer : Nairita GHOSAL, Merck & Co., INC.

16:15 Nonparametric Bayesian Analysis of Genotoxicity Screening Assays [Abstract 97]

Dingzhou (Dean) LI, *Drug Safety Statistics, Pfizer,Drug Safety Statistics, Pfizer*

16:40 Semiparametric Bayesian Markov Analysis of Personalized Benefit-Risk Assessment [Abstract 197]

Dongyan YAN, *Discovery & Development Statistics,Eli Lilly and Company*
Subharup GUHA, *Department of Biostatistics, University of Florida*
Chul AHN, *Division of Biostatistics, Center for Devices and Radiological Health, Office Surveillance and Biometrics, U.S. Food and Drug Administration*
Ram TIWARI, *Statistical Methodology at Bristol Myers Squibb*

17:05 Clustered-Temporal Bayesian Model for Brain connectivity in Neuroimaging data [Abstract 53]

Nairita GHOSAL, *Merck & Co., INC.*
Sanjib BASU, *University of Illinois at Chicago*

IS 53 Statistical Methods for Contemporary Applications *Venue:* Room 5

Chair : Trambak BANERJEE, Analytics, Information and Operations Management,University of Kansas

Organizer : Gourab MUKHERJEE, Department of Data Sciences & Operations,University of Southern California

16:15 **A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data [Abstract 8]**

Trambak BANERJEE, *Analytics, Information and Operations Management,University of Kansas*
Bhaswar BHATTACHARYA, *University of Pennsylvania*
Gourab MUKHERJEE, *University of Southern California*

16:40 **Applying Double Machine Learning to Targeted Email Promotions: A Journey Down the Conversion Funnel [Abstract 79]**

Wreetabrata KAR, *Purdue University,Purdue University-Main Campus (West Lafayette, IN)*

17:05 **A regression modeling approach to simultaneous estimation [Abstract 210]**

Dave ZHAO, *Statistics,University of Illinois at Urbana-Champaign*

IS 54 Topics in Network Inference *Venue:* Room 6

Chair : Debashis MONDAL, Oregon State University

Organizer : Sharmodeep BHATTACHARYYA, Oregon State University

16:15 **Randomization-only Inference in Experiments with Interference [Abstract 34]**

David CHOI, *Heinz College,Carnegie Mellon University*

16:40 **Bias-Variance Tradeoffs in Joint Spectral Embeddings [Abstract 178]**

Daniel SUSSMAN, *Mathematics and Statistics,Boston University*
Benjamin DRAVES, *Boston University*

17:05 **Linear regression and its inference on noisy network-linked data [Abstract 98]**

Tianxi LI, *Department of Statistics,University of Virginia*
Can M. LE, *UC Davis*

Saturday May 22

Plenary Lecture 3 Xihong Lin

Venue: Room E

Chair : Susmita DATTA, University of Florida

8:30 **Scalable Integrative Analysis of Large Genome and Phenome Data** [**Abstract** 105]

Xihong LIN, *Departments of Biostatistics and Department of Statistics, Harvard University and Broad Institute*

Special Invited Session 3 Roshan Joseph and Bani Kumar Mallick

Venue: Room E

Chair : Sujit GHOSH, Department of Statistics,North Carolina State University

9:45 **Data Splitting** [**Abstract** 76]

Roshan JOSEPH, *School of Industrial and Systems Engineering,Georgia Institute of Technology*
Akhil VAKAYIL, *Georgia Institute of Technology*

10:30 **Bayesian local models using partitions** [**Abstract** 116]

Bani MALLICK, *Statistics,Texas A&M*

Special Invited Session 4 Frank Bretz and Aloka Chakravarty

Venue: Room F

Chair : Shanthi SETHURAMAN, Eli Lilly and Company

9:45 **Equivalence of regression curves** [**Abstract** 18]

Frank BRETZ, *Statistical Methodology,Novartis Pharma AG*
Kathirn MOELLENHOFF, *Eindhoven University of Technology*
Holger DETTE, *University of Bochum*

10:30 **Generating Actionable Insights from Real World Data during COVID-19 Pandemic** [**Abstract** 24]

Aloka CHAKRAVARTY, *Immediate Office of the Commissioner, FDA,US Food and Drug Administration*

IS 55 Biomarker Characterization and Development for Patient Enrichment in Oncology Clinical Trial

Venue: Room 1

Chair : Pralay MUKHOPADHYAY, Department of Biometrics,Otsuka America Pharmaceuticals Inc.

Organizer : Feng LIU, Oncology Biometrics,AstraZeneca

11:30 **Statistical Considerations in Biomarker Cutoff Determination, Development and Validation in Immuno-Oncology Studies** [**Abstract** 201]

Jiabu YE, *Oncology Biometrics AstraZeneca,AstraZeneca*
Feng LIU, *Oncology Biometrics AstraZeneca*

11:55 **Blocked adaptive randomization: how to do RAR if you really must.** [**Abstract** 25]

Richard CHAPPELL, *U Wisconsin Dept. of Biostatistics and Medical Informatics,University of Wisconsin*
Thevaasinen CHANDERENG, *Columbia University*

12:20 **Design and Challenges in Platform Design in the Immuno-Oncology Drug Development with application in NSCLC** [**Abstract** 107]

Feng LIU, *Oncology Biometrics,AstraZeneca*

IS 56 Radomization and More

Venue: Room 2

Chair : Jie YANG, Department of Mathematics, Statistics, and Computer Science,University of Illinois at Chicago

Organizer : Abhyuday MANDAL, University of Georgia

11:30 Randomization tests of causal effects under general interference [Abstract 145]

David PUELZ, *Booth School of Business,University of Chicago*

11:55 Optimal Crossover Designs for Generalized Linear Models [Abstract 70]

Jeevan JANKAR, *Department of Statistics,University of Georgia, Athens*
Abhyuday MANDAL, *University of Georgia, Athens*
Jie YANG, *University of Illinois, Chicago*

12:20 Rerandomization and Regression Adjustment [Abstract 101]

Xinran LI, *Department of Statistics,University of Illinois at Urbana-Champaign*
Peng DING, *University of California, Berkeley*

IS 57 Subsampling Methods in the Era of Big Data

Venue: Room 3

Chair : John STUFKEN, Informatics and Analytics,University of North Carolina at Greensboro

Organizer : Sujit GHOSH, Department of Statistics,North Carolina State University

11:30 Supervised compression of data [Abstract 115]

Simon MAK, *Statistical Science,Duke University*
V. Roshan JOSEPH, *Georgia Institute of Technology*

11:55 Sampling for Massive Data with Rare Events [Abstract 187]

Hai Ying WANG, *Statistics,University of Connecticut*

12:20 Generalized orthogonal subsampling for predictive stability [Abstract 190]

Lin WANG, *Department of Statistics,George Washington University*
Yi ZHANG, *George Washington University*

IS 58 Contemporary Bayesian Modeling with Applications

Venue: Room 4

Chair : Aritra HALDER, gxk9jg@virginia.edu

Organizer : Dipak DEY, Department of Statistics,University of Connecticut

11:30 Dependent Mixtures: Modeling Cell Lineage [Abstract 126]

Peter MUELLER, *Statistics & Data Sc,UT Austin*
Carlos Pagani ZANINI, *UFRJ*
Giorgio PAULON, *UT Austin*

11:55 Some recent models using binary tree ensembles for various outcome types [Abstract 94]

Purushottam LAUD, *Medical College of Wisconsin*
Robert MCCULLOCH, *Arizona State University*
Rodney SPARAPANI, *Medical College of Wisconsin*
Brent LOGAN, *Medical College of Wisconsin*

12:20 An Objective Bayesian Multiple Testing of Binomial Proportions [Abstract 169]

Siva SIVAGANESAN, *Division of Statistics and Data Science, Department of Mathemtical Sciences,University of Cincinnati*
Emrah GECILI, *Cincinnati Children's Hospital*
Nilupika HERATH, *University of Cincinnati*

IS 59 Advances in Random Forests and Decision Tree Ensembles *Venue:* Room 5

Chair and Organizer : Sumanta BASU, Statistics and Data Science,Cornell University

11:30 Random Forests: Why They Work and Why That's a Problem [Abstract 123]

Lucas MENTCH, *Department of Statistics,University of Pittsburgh*
Siyu ZHOU, *University of Pittsburgh*

11:55 Improved uncertainty quantification for random forests and other ensembles [Abstract 64]

Giles HOOKER, *Statistics and Data Science,Cornel University*
Zhengze ZHOU, *Cornell University*
Indrayudh GHOSAL, *Cornell University*

12:20 Nonparametric Variable Screening with Decision Stumps [Abstract 85]

Jason KLUSOWSKI, *Operations Research and Financial Engineering,Princeton University*
Peter M. TIAN, *Princeton University*

IS 60 Advances in Analysis of Intensive Longitudinal Data *Venue:* Room 6

Chair and Organizer : Donald HEDEKER, University of Chicago

11:30 Mixed-effects location scale modeling for the analysis of accelerometry data [Abstract 194]

Whitney WELCH, *Department of Preventive Medicine,Northwestern University Feinberg School of Medicine*
Donald HEDEKER, *University of Chicago*
Bonnie SPRING, *Northwestern University Feinberg School of Medicine*
Juned SIDDIQUE, *Northwestern University Feinberg School of Medicine*

11:55 A Negative Binomial Mixed Effects Location-Scale Model for Physical Activity Data Provided by Wearable Devices [Abstract 110]

Qianheng MA, *Department of Public Health Sciences,University of Chicago*
Genevieve F. DUNTON, *University of Southern California*
Donald HEDEKER, *University of Chicago*

12:20 Latent Trait Shared Parameter Mixed Models for Missing Ordinal Ecological Momentary Assessment Data [Abstract 39]

John CURSIO, *Public Health Sciences,University of Chicago*

IS 61 Novel Methods for Oncology Drug Development *Venue:* Room 1

Chair and Organizer : Arnab MAITY, Pfizer

13:00 Bayesian Approach for Single-Arm Trials Incorporating Historical Information [Abstract 9]

Cynthia BASU, *Early Clinical Development,Pfizer Inc.*
Arnab K. MAITY, *Pfizer Inc.*
Lada A. MARKOVTSOVA, *Pfizer Inc.*
Satrajit ROYCHOUDHURY, *Pfizer Inc.*

13:25 Model based methods to assess the impact of missing data on Oncology endpoints [Abstract 23]

Arunava CHAKRAVARTTY, *Biostatistics/Novartis,Novartis*
Craig WANG, *Novartis*

13:50 On weighted log-rank combination tests and companion Cox model estimators [**Abstract 96**]

Larry LEON, *Biostatistics,Bristol-Myers Squibb*
Ray LIN, *Genentech*
Keaven ANDERSON, *Merck*

IS 62 Efficient Designs *Venue:* Room 2

Chair : John STUFKEN, Informatics and Analytics,University of North Carolina at Greensboro
Organizer : Abhyuday MANDAL, University of Georgia

13:00 Selection of 2-level supersaturated designs for main effects models [**Abstract 167**]

Rakhi SINGH, *Informatics and Analytics,University of North Carolina at Greensboro*
John STUFKEN, *University of North Carolina at Greensboro*

13:25 Min–Max Crossover Designs for Two Treatments Binary and Poisson Crossover Trials [**Abstract 168**]

Satya SINGH, *Mathematics,Indian Institute of Technology Hyderabad*
Siuli MUKHOPADHYAY, *Indian Institute of Technology of Bombay*
Harsh RAJ, *Indian Institute of Technology of Hyderabad*

13:50 Fast Approximation of Shapley Values [**Abstract 214**]

Wei ZHENG, *Business Analytics and Statistics,university of tennessee*
liuqing YANG, *Nankai University*
Yongdao ZHOU, *Nankai University*
Haoda FU, *Eli Lilly and Company*
Minqian LIU, *Nankai University*

IS 63 Modeling and Analysis of High Dimensional Graphs, Networks, and Tensors *Venue:* Room 3

Chair and Organizer : Naveen Naidu NARISETTY, Statistics,University of Illinois at Urbana-Champaign

13:00 Edgeworth expansion for network moments [**Abstract 209**]

Yuan ZHANG, *Statistics,The Ohio State University*
Dong XIA, *Hong Kong University of Science and Technology*

13:25 An Optimal Statistical and Computational Framework for Generalized Tensor Estimation [**Abstract 205**]

Anru ZHANG, *Statistics,University of Wisconsin-Madison / Duke University*

13:50 Bayesian Regularization and Estimation for Gaussian Conditional Random Fields [**Abstract 102**]

Feng LIANG, *Statistics Department, UIUC,University of Illinois at Urbana-Champaign*
Lingrui GAN, *Facebook*
Naveen N. NARISETTY, *University of Illinois at Urbana-Champaign*

IS 64 Advances in High-Dimensional Bayesian Methodology *Venue:* Room 4

Chair and Organizer : Kshitij KHARE, Department of Statistics, University of Florida

13:00 Dynamics of mean-field approximation: a case-study in singular models [**Abstract 13**]

Anirban BHATTACHARYA, *Statistics,Texas A&M University*
Debdeep PATI, *Texas A&M University*
Yun YANG, *University of Illinois at Urbana-Champaign*
Sean PLUMMER, *Texas A&M University*

13:25 **Statistical optimality and stability of tangent transform algorithms** [**Abstract 139**]

Debdeep PATI, *Statistics,Texas A&M University*
Anirban BHATTACHARYA, *Texas A&M University*
Indrajit GHOSH, *Texas A&M University*

13:50 **Lag selection and estimation in mixed frequency regression using Bayesian nested lasso** [**Abstract 83**]

Kshitij KHARE, *Department of Statistics, University of Florida*
Satyajit GHOSH, *FDA*
George MICHAILIDIS, *University of Florida*

IS 65 Statistical methods for genetic prediction of disease using 'omics data *Venue:* Room 5

Chair and Organizer : Brandon COOMBES, Department of Quantitative Health Sciences, Mayo Clinic

13:00 **Application of polygenic risk scores to diverse populations** [**Abstract 38**]

Brandon COOMBES, *Department of Quantitative Health Sciences, Mayo Clinic*
Anthony BATZER, *Mayo Clinic*
Gregory JENKINS, *Mayo Clinic*
Euijung RYU, *Mayo Clinic*

13:25 **Novel strategy for disease risk prediction incorporating predicted gene expression and DNA methylation: a multi-phased study of prostate cancer** [**Abstract 195**]

Chong WU, *Department of Statistics, Florida State University,Department of Statistics, Florida State University*
Jingjing ZHU, *Cancer Epidemiology Division, Population Sciences in the Pacific Program, University of Hawaii Cancer Center, University of Hawaii at Manoa, Honolulu, HI, USA*
Austin KING, *Department of Statistics, Florida State University*
Xiaoran TONG, *Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA*
Qing Lu, Jong Y Park, Liang Wang, Guimin Gao, Hong-Wen Deng, Yaohua Yang, Karen E Knudsen, Timothy R Rebbeck, Jirong Long, Wei Zheng, Wei Pan, David V Conti, Christopher A Haiman, Lang WU,

13:50 **Discussant: Brandon COOMBES**, Department of Quantitative Health Sciences, Mayo Clinic

IS 66 Modern computational methods for spatial data *Venue:* Room 6

Chair and Organizer : Abhirup DATTA, Johns Hopkins University

13:00 **A matrix-free profile likelihood method for high-dimensional factor model** [**Abstract 46**]

Somak DUTTA, *Statistics,Iowa State University*
Fan DAI, *Michigan Technological University*
Ranjan MAITRA, *Iowa State University*

13:25 **A Joint Spatial Conditional Auto-Regressive Model for Estimating HIV Prevalence Rates Among Key Populations** [**Abstract 93**]

Zhou LAN, *Yale School of Medicine,Yale School of Medicine*
Le BAO, *Penn State University*

13:50 **Test for Isotropy on a Sphere using Spherical Harmonic Coefficients** [**Abstract 157**]

Indranil SAHOO, *Statistical Sciences & Operations Research, Virginia Commonwealth University*
Joseph GUINNESS, *Cornell University*
Brian J. REICH, *North Carolina State University*

Sunday May 23

In Memorium: Professor Hukum Chandra (1972 - 2021)

Venue: Room GB

Time : **10:00 - 10:30 CDT**

A registration, separate from the IISA 2021 is required. The registration is free. Please register here. The system will send the necessary links to the e-mail address used for registration.

Break 10:30 - 10:40 CDT

Meeting 1 IISA General Body Meeting

Venue: Room GB

Time : **10:40 - 12:00 CDT**

- IISA General Body Meeting

  The room could be accessed from the same link as above.

Abstracts

1. Adjusted chi-square test for degree-corrected block models

[IS 31, (page 15)]

Arash AMINI, *Statistics, UCLA*
Linfan ZHANG, *UCLA*

We propose a goodness-of-fit test for degree-corrected stochastic block models (DCSBM). The test is based on an adjusted chi-square statistic for measuring equality of means among groups of n multinomial distributions with $d_1, \dots, d_n$ observations. In the context of network models, the number of multinomials, n, grows much faster than the number of observations, d_i, corresponding to the degree of node i, hence the setting deviates from classical asymptotics. We show that a simple adjustment allows the statistic to converge in distribution, under null, as long as the harmonic mean of $\{d_i\}$ grows to infinity.

When applied sequentially, the test can also be used to determine the number of communities. The test operates on a (row) compressed version of the adjacency matrix, conditional on the degrees, and as a result is highly scalable to large sparse networks. We incorporate a novel idea of compressing the rows based on a $(K+1)$-community assignment when testing for K communities. This approach increases the power in sequential applications without sacrificing computational efficiency, and we prove its consistency in recovering the number of communities. Since the test statistic does not rely on a specific alternative, its utility goes beyond sequential testing and can be used to simultaneously test against a wide range of alternatives outside the DCSBM family.

The test can also be easily applied to Poisson count arrays in clustering or biclustering applications, as well as bipartite and directed networks. We show the effectiveness of the approach by extensive numerical experiments with simulated and real data. In particular, applying the test to the Facebook-100 dataset, a collection of one hundred social networks, we find that a DCSBM with a small number of communities (say < 25) is far from a good fit in almost all cases. Despite the lack of fit, we show that the statistic itself can be used as an effective tool for exploring community structure, allowing us to construct a community profile for each network.

2 . ODE backpropagation dynamics and reparameterization for high dimensional lensing inference from the CMB polarization

[IS 9, (page 6)]

Ethan ANDERES, *Department of Statistics, University of California at Davis*

In the last decade cosmologists have spent a considerable amount of effort mapping the radially-projected large-scale mass distribution in the universe by measuring the distortion it imprints on the CMB. Indeed, all the major surveys of the CMB produce estimated maps of the projected gravitational potential generated by mass density fluctuations over the sky. These maps contain a wealth of cosmological information and, as such, are an important data product of CMB experiments. However, the most profound impact from CMB lensing studies may not come from measuring the lensing effect, per se, but rather from our ability to remove it, a process called delensing. This is due to the fact that lensing, along with emission of millimeter wavelength radiation from the interstellar medium in our own galaxy, are the two dominant sources of foreground contaminants for primordial gravitational wave signals in the CMB polarization. As such delensing, i.e. the process of removing the lensing contaminants, and our ability to either model or remove galactic foreground emission sets the noise floor on upcoming gravitational wave science.

In this talk we will present a complete Bayesian solution for simultaneous inference of lensing, delensing and gravitational wave signals in the CMB polarization as characterized by the tensor-to-scalar ratio r parameter. Our solution relies crucially on a physically motivated re-parameterization of the CMB polarization which is designed specifically, along with the design of the Gibbs Markov chain itself, to result in an efficient Gibbs sampler—in terms of mixing time and the computational cost of each step—of the Bayesian posterior. This re-parameterization also takes advantage of a newly developed lensing algorithm, which we term LenseFlow, that lenses a map by solving a system of ordinary differential equations. This description has conceptual advantages, such as allowing us to give a simple non-perturbative proof that the lensing determinant is equal to unity in the weak-lensing regime. The algorithm itself maintains this property even on pixelized maps, which is crucial for our purposes and unique to LenseFlow as compared to other lensing algorithms we have tested. It also has other useful properties such as that it can be trivially inverted (i.e. delensing) for the same computational cost as the forward operation, and can be used for fast and exact likelihood gradients with respect to the lensing potential. Incidentally, the ODEs for calculating these derivatives are exactly analogous to the backpropagation techniques used in deep neural networks but are derived in this case completely from ODE theory.

3. Do MaxCombo and Weighted Logrank Tests Control Type I Error?

[IS 44, (page 20)]

Keaven ANDERSON, *Methodology Research,Merck & Co., Inc.*
Pralay MUKHOPADHYAY, *Department of Biometrics,Otsuka America Pharmaceuticals Inc.*
Satrajit ROYCHOUDHURY, *Pfizer inc.*

A pharmaceutical industry working group has proposed a combination test of Fleming-Harrington weighted logrank tests to control Type I error and ensure robust power across a range of alternative hypotheses. These tests have recently been criticized in the literature for not adequately controlling Type I error. We review some currently noted weighted logrank tests and examine these concerns. In general, we have found the criticisms not to represent concerns. However, we also propose alternatives to the originally proposed tests and discuss their application.

4. Bayesian Integrative Approaches to Enable Precision Medicine

[IS 25, (page 12)]

Veera BALADANDAYUTHAPANI, *Biostatistics,University of Michigan*
Min Jin Ha, Abhisek Saha, Satwik ACHARYYA, *MD Anderson, NIH, University of Michigan*

At the heart of Precision Medicine is connecting the right drug/therapy to the right patient. The extensive acquisition of high-throughput molecular and drug profiling data across diverse model systems have made precision medicine efforts a realistic possibility. Modern precision medicine endeavors are at an inflexion point facing the fundamental challenge of assimilating, organizing, analyzing and interpreting multi-domain data types to make individualized health decisions. From an analytic viewpoint, modeling and inference in such studies is challenging, not only due to high dimensionality, but also due to presence of structured dependencies (e.g. pathway/regulatory mechanisms, serial and spatial correlations). Integrative analyses of these multi-domain data, combined with patients clinical outcomes, can help us quantify and interpret the complex biological processes that characterize a disease. This talk will cover probabilistic Bayesian statistical and computational frameworks that acknowledge and exploit these inherent complex structural relationships, for both biomarker discovery, and clinical prediction, to aid evidence-based translational and individualized medicine.

5 . Adaptive and powerful microbiome multivariate association analysis via feature selection

[Student Paper Competition 2, (page 19)]

Kalins BANERJEE, *Department of Public Health Sciences,Pennsylvania State University*
Jun CHEN, *Division of Biomedical Statistics and Informatics, Mayo Clinic*
Xiang ZHAN, *Department of Public Health Sciences, Pennsylvania State University*

The importance of human microbiome is being increasingly recognized, and researchers are more than ever interested in studying associations between microbial features and human health or disease conditions. Microbiome data are high dimensional, and some or even most microbial taxa often might not be associated with the outcome of interest. Realizing the existing methods' susceptibility to the adverse effects of noise accumulation, we introduce Adaptive Microbiome Association Test (AMAT), a novel and powerful tool for multivariate microbiome association analysis, which unifies the blessings of feature selection in high-dimensional multivariate inference and the robustness of adaptive tests. AMAT first alleviates the burden of noise accumulation via distance correlation learning, and then conducts a data-adaptive association test under the flexible generalized linear model framework. Extensive simulation studies and real data applications demonstrate that AMAT is highly robust and often more powerful than several existing methods, while preserving the correct type I error rate. Besides, applying AMAT to a high taxonomic rank can provide a useful preliminary screening, and facilitate more targeted downstream microbiome fine mapping.

6 . Statistical Issues and Challenges in Analyzing Data from a Quality Improvement Collaborative

[Special Invited Session 2, (page 7)]

Mousumi BANERJEE, *Biostatistics,University of Michigan*

The Pediatric Cardiac Critical Care Consortium (PC^4) is a quality collaborative that aims to improve the quality of care to patients with critical pediatric and congenital cardiovascular disease in North America and abroad. Formed in 2009 with National Institutes of Health funding, PC^4 is a unique collaborative of leaders in pediatric cardiac critical care, cardiac surgery, and cardiology representing a diverse group of centers caring for these vulnerable patients. The core pillars of collaborative quality improvement serve as the foundation for PC^4: purposeful collection of specific clinical data on outcomes and practice; timely performance feedback to clinicians, and continuous improvement based on empirical analysis and collaborative learning. This talk will provide an overview of the statistical issues and challenges in analyzing data from a quality improvement collaborative based on my experience as PC^4 statistician. Specifically, I will discuss examples from three broad focus areas of the collaborative, namely to 1) develop clinical decision making tools to predict outcomes in pediatric cardiac intensive care units (PCICU) using multiplatform data, 2) quantify variation in risk-adjusted outcomes across PCICUs after surgery, and benchmark performance, and 3) examine organizational and personnel factors as they may relate to post-surgical outcomes. This is an applications oriented talk that will describe novel applications of predictive modeling and causal inference in a quality collaborative setting.

7. Minimum Cost-Compression Risk in Principal Component Analysis

[IS 39, (page 18)]

Swarnali BANERJEE, *Department of Mathematics and Statistics, Loyola University Chicago*
Bhargab CHATTOPADHYAY, *Indian Institute of Management Visakhapatnam*

Principal Component Analysis (PCA) is commonly used to reduce dimension of data instances into its subspace retaining a subset of linearly uncorrelated variables or principal components. With high-dimensional datasets, there is a need to perform PCA on data sets as soon as they are observed and results are updated without computing from the beginning. Online PCA may be used for time varying data sets and also for ones that are too large and can be processed in batches. Minimizing expected quadratic compression loss requires a large number of observations thus rendering the process expensive and time consuming. We consider a cost-compression loss which takes into account the data compression loss as well as the cost of collecting data and minimizes the corresponding risk. We propose a fast two-stage algorithm that enjoys convergence properties and as an extension, a purely sequential algorithm with no assumption on specific data distribution. The performance is analyzed through applications in a financial data set.

8. A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data

[IS 53, (page 24)]

Trambak BANERJEE, *Analytics, Information and Operations Management,University of Kansas*
Bhaswar BHATTACHARYA, *University of Pennsylvania*
Gourab MUKHERJEE, *University of Southern California*

An important problem in contemporary immunology studies based on single-cell protein expression data is to determine whether cellular expressions are remodeled post infection by a pathogen. One natural approach for detecting such changes is to use nonparametric two-sample statistical tests. However, in single-cell studies, direct application of these tests is often inadequate, because single-cell level expression data from processed uninfected populations often contain attributes of several latent subpopulations with highly heterogeneous characteristics. As a result, viruses often infect these different subpopulations at different rates in which case the traditional nonparametric two-sample tests for checking similarity in distributions are no longer conservative. In this paper, we propose a new nonparametric method for Testing Remodeling Under Heterogeneity (TRUH) that can accurately detect changes in the infected samples compared to possibly heterogeneous uninfected samples. Our testing framework is based on composite nulls and is designed to allow the null model to encompass the possibility that the infected samples, though unaltered by the virus, might be dominantly arising from under-represented subpopulations in the baseline data. The TRUH statistic, which uses nearest neighbor projections of the infected samples into the baseline uninfected population, is calibrated using a novel bootstrap algorithm. We demonstrate the non-asymptotic performance of the test via simulation experiments, and also derive the large sample limit of the test statistic, which provides theoretical support towards consistent asymptotic calibration of the test. We use the TRUH statistic for studying remodeling in tonsillar T cells under different types of HIV infection and find that unlike traditional tests which do not have any heterogeneity correction, TRUH based statistical inference conforms to the biologically validated immunological theories on HIV infection.

9. Bayesian Approach for Single-Arm Trials Incorporating Historical Information

[IS 61, (page 27)]

Cynthia BASU, *Early Clinical Development,Pfizer Inc.*
Arnab K. MAITY, *Pfizer Inc.*
Lada A. MARKOVTSOVA, *Pfizer Inc.*
Satrajit ROYCHOUDHURY, *Pfizer Inc.*

Interim analysis is an important key in clinical trials to inform the ongoing trial about the potential success (or failure) at the of the study to allow for optimal resource allocation and patient safety. In this presentation we focus on some Bayesian approaches for single arm trials to compute such success probabilities and to make decision about go-no go. Moreover, we discuss borrowing historical information in carrying out these analyses. We demonstrate the methods for two endpoints binary and time to event. In particular, when analyzing the time to event endpoints we provide simple analytical expressions for predictive probability which is essential to calculate the interim probabilities, and thus we are able to bypass the Markov Chain Monte Caro (MCMC) sampling.

10. Efficient SNP-based Heritability Estimation using Gaussian Predictive Process in Large-scale Cohort Studies

[IS 37, (page 17)]

Saonli BASU, *Division of Biostatistics, University of Minnesota*
Souvik SEAL, *Colorado School of Public Health*
Abhirup DATTA, *Johns Hopkins University*

For decades, linear mixed models (LMM) have been widely used to estimate heritability in twin and family studies. Recently, with the advent of high throughput genetic data, there have been attempts to estimate heritability from genome-wide SNP data on a cohort of distantly related individuals. Fitting such an LMM in large-scale cohort studies, however, is tremendously challenging due to high dimensional linear algebraic operations. In this paper, we simplify the LMM by unifying the concept of Genetic Coalescence and Gaussian Predictive Process, and thereby greatly alleviating the computational burden. Our proposed approach PredLMM has much better computational complexity than most of the existing packages and thus, provides an efficient alternative for estimating heritability in large-scale cohort studies. We illustrate our approach with extensive simulation studies and use it to estimate the heritability of multiple quantitative traits from the UK Biobank cohort.

11. Large Spectral Density Matrix Estimation by Thresholding

[IS 11, (page 6)]

Sumanta BASU, *Statistics and Data Science,Cornell Uniersity*

Spectral density matrix estimation of multivariate time series is a classical problem in time series and signal processing. In modern neuroscience, spectral density based metrics are commonly used for analyzing functional connectivity among brain regions. In this paper, we develop a non-asymptotic theory for regularized estimation of high-dimensional spectral density matrices of Gaussian and linear processes using thresholded versions of averaged periodograms. Our theoretical analysis ensures that consistent estimation of the spectral density matrix of a p-dimensional time series using n samples is possible under high-dimensional regime log p = o(n) as long as the true spectral density is approximately sparse. A key technical component of our analysis is a new concentration inequality of average periodogram around its expectation, which is of independent interest. Our estimation consistency results complement existing results for shrinkage based estimators of multivariate spectral density, which require no assumption on sparsity but only ensure consistent estimation in a regime $p^2 = o(n)$. In addition, our proposed thresholding based estimators perform consistent and automatic edge selection when learning coherence networks among the components of a multivariate time series. We demonstrate the advantage of our estimators using simulation studies and a real data application on functional connectivity analysis with fMRI data.

12. Edge Selection for Graphical Models with Mixed Types under Informative Sampling

[IS 10, (page 6)]

Emily BERG, *Statistics,Iowa State University*
Hao SUN, *Iowa State University*
Zhengyuan ZHU, *Iowa State University*

Complex surveys often collect high-dimensional vectors composed of discrete and continuous variables. Exploratory questions of interest are about multivariate associations in the data. We propose to use graphical models to analyze high-dimensional survey data. We view the survey questions as nodes and use the edges to specify relationships between the nodes. The variables at the nodes of the graph may be discrete or continuous. The graphical representation defines an unknown joint distribution for the superpopulation model. We observe a sample selected from the superpopulation. Edge selection methods that assume simple random sampling can lead to biased estimates of the graph structure. We investigate edge selection methods for complex survey designs.

13. Dynamics of mean-field approximation: a case-study in singular models

[IS 64, (page 28)]

Anirban BHATTACHARYA, *Statistics,Texas A&M University*
Debdeep PATI, *Texas A&M University*
Yun YANG, *University of Illinois at Urbana-Champaign*
Sean PLUMMER, *Texas A&M University*

The marginal likelihood or evidence in Bayesian statistics contains an intrinsic penalty for larger model sizes and is a fundamental quantity in Bayesian model comparison. Over the past two decades, there has been steadily increasing activity to understand the nature of this penalty in singular statistical models, building on pioneering work by Sumio Watanabe. Unlike regular models where the Bayesian information criterion (BIC) encapsulates a first-order expansion of the logarithm of the marginal likelihood, parameter counting gets trickier in singular models where a quantity called the real log canonical threshold (RLCT) summarizes the effective model dimensionality. We show that mean-field variational inference correctly recovers the RLCT for any singular model in its canonical or normal form. We additionally exhibit sharpness of our bound by analyzing the dynamics of a general purpose coordinate ascent algorithm (CAVI) popularly employed in variational inference.

14 . Motif Estimation via Subgraph Sampling: The Fourth-Moment Phenomenon

[IS 17, (page 9)]

Bhaswar BHATTACHARYA, *Statistics,University of Pennsylvania*
Sayan DAS, *Columbia University*
Sumit MUKHERJEE, *Columbia University*

Network sampling has emerged as an indispensable tool for understanding features of large-scale complex networks where it is practically impossible to search/query over all the nodes. Examples include social networks, biological networks, internet and communication networks, and socio-economic networks, among others. In this talk we will discuss a unified framework for statistical inference for counting motifs, such as edges, triangles, and wedges, in the widely used subgraph sampling model. In particular, we will provide precise conditions for the consistency and the asymptotic normality of the natural HorvitzThompson (HT) estimator, which can be used for constructing confidence intervals and hypothesis testing for the motif counts. As a consequence, an interesting fourth-moment phenomena for the asymptotic normality of the HT estimator and connections to fundamental results in random graph theory will emerge.

15. Assessing contribution of treatment phases through tipping point analyses using rank preserving structural failure time models

[IS 27, (page 13)]

Sudipta BHATTACHARYA, *SQS, Biostatistics,Takeda*
Jyotirmoy DEY, *Regeneron*

The proposed work provides a novel approach for assessing the importance of specific treatment phases in clinical research through tipping point analyses (TPA) of a time-to-event endpoint using rank-preserving-structural-failure-time (RPSFT) modeling. In oncology clinical research, an experimental treatment is often added to standard of care therapy in multiple treatment phases (e.g., concomitant and maintenance phases) to improve patient outcomes. When the resulting new regimen provides meaningful benefit over standard of care, gaining insights into the contribution of each treatment phase becomes important to properly guide clinical practice. Since traditional statistical approaches are inadequate for answering such questions, new approaches are needed. RPSFT modeling is an approach for causal inference, typically used to adjust for treatment switching in randomized clinical trials with time to event endpoints. A tipping-point analysis is commonly used in situations where a statistically significant treatment effect is suspected to be an artifact of missing or unobserved data rather than a real treatment effect. The methodology proposed here is an amalgamation of these two ideas to investigate the contribution of a specific component of a regimen comprising multiple treatment phases.

16. TransPRECISE: Personalized Network Modeling of the Pan-cancer Patient and Cell Line Interactome

[IS 23, (page 11)]

Rupam BHATTACHARYYA, *Biostatistics,University of Michigan*
Min Jin HA, *The University of Texas MD Anderson Cancer Center*
Qingzhi LIU, *University of Michigan*
Rehan AKBANI, *The University of Texas MD Anderson Cancer Center*
Han LIANG, *The University of Texas MD Anderson Cancer Center*

A systemic approach in precision medicine has been to bridge anticancer pharmacological data to large-scale multi-layered molecular tumor profiles using cell lines as proxies for cancer patients. However, samples from different tumor microenvironments in the two model systems may exhibit distinct patterns of molecular activities in general. Specifically, the architecture of cancer modulation through cumulative effects from multiple interacting genes in functional/signaling pathways may vary across model systems. In this work, we attempt to address these challenges by developing a multi-level Bayesian analytical framework called TransPRECISE: Translational Personalized Cancer-specific Integrated Network Estimation. TransPRECISE uses Bayesian graphical regression models to infer on cancer-specific pathway circuitry by establishing weighted connections between the genes in the pathway based on the posterior inclusion probabilities of the protein-protein interactions. These cancer-specific networks can then be deconvolved to a sample-specific level by identifying node (gene/protein) labels as neutral, activated or suppressed, based on the extremity (both in magnitude and direction) of each sample-specific gene expression with respect to the fitted model. These sample-specific networks can further be summarized to neutrality, activation or suppression scores and statuses of each pathway for each sample. We use TransPRECISE to analyze proteomic data for 640 cell lines from 16 lineages and 7714 patients from 31 tumor types. Through pan-cancer analysis, we investigate differential and conserved aspects of cancer-specific pathway networks across model systems and cancer lineages. Additionally, we identify matching avatar cell lines for patient tumor profiles via correlating the sample-specific scores across patient and cell line cancer types, based on a pan-pathway approach using Pearson correlations and another pathway-specific approach using hierarchical clustering. We also train Bayesian additive regression tree (BART) models based on cell lines for predicting drug sensitivity in patients. Finally, we developed an R Shiny app for aiding seamless visualization of results and possible future investigations into our findings.

17. Detecting rare haplotype association with two correlated phenotypes of binary and continuous types

[IS 48, (page 22)]

Swati BISWAS, *Mathematical Sciences,University of Texas at Dallas*
Xiaochen YUAN, *University of Texas at Dallas*

Multiple correlated traits/phenotypes are often collected in genetic association studies and they may share a common genetic mechanism. Joint analysis of correlated phenotypes has well known advantages over one-at-a-time analysis including gain in power and better understanding of genetic etiology. However, when the phenotypes are of discordant types such as binary and continuous, the joint modeling is more challenging. Another research area of current interest is discovery of rare genetic variants. Currently there is no method available for detecting association of rare (or common) haplotypes with multiple discordant phenotypes jointly. Our goal is to fill this gap specifically for two discordant phenotypes. We consider a rare haplotype association method for a binary phenotype Logistic Bayesian LASSO (univariate LBL) and its extension for two correlated binary phenotypes (bivariate LBL-2B). Under this framework, we propose a haplotype association test with binary and continuous phenotypes jointly (bivariate LBL-BC). Specifically, we use a latent variable to induce correlation between the two phenotypes. We carry out extensive simulations to investigate bivariate LBL-BC and compare it with univariate LBL and bivariate LBL-2B. In most settings, bivariate LBL-BC performs the best. In only two situations, bivariate LBL-BC has similar performance - when the two phenotypes are (1) weakly or not correlated and the target haplotype affects the binary phenotype only and (2) strongly positively correlated and the target haplotype affects both phenotypes in positive direction. Finally, we apply the method to a dataset on lung cancer and nicotine dependence and detect several haplotypes including a rare one.

18. Equivalence of regression curves

[Special Invited Session 4, (page 25)]

Frank BRETZ, *Statistical Methodology,Novartis Pharma AG*
Kathirn MOELLENHOFF, *Eindhoven University of Technology*
Holger DETTE, *University of Bochum*

We investigate the problem whether the difference between two parametric models describing the relation between a response variable and several covariates in two non-overlapping populations is practically irrelevant, such that inference can be performed on the basis of the pooled sample. Statistical methodology is developed to demonstrate equivalence between the two regression curves for a pre-specified equivalence margin. We extend this methodology to other situations of practical interest, such as to establish the equivalence of two regression curves for correlated binary outcomes or when the regression models share common parameters. We illustrate the proposed methodology by means of clinical trial examples and investigate the operating characteristics through an extensive simulation study.

References: [1] Dette H, Moellenhoff K, Volgushev S, Bretz F (2018) Equivalence of regression curves. Journal of the American Statistical Association 113(522), 711-729. [2] Moellenhoff K, Bretz F, Dette H. (2020) Equivalence of regression curves sharing common parameters. Biometrics 76, 518529. [3] Moellenhoff K, Dette H, Bretz F (2021) Testing for similarity of binary efficacy-toxicity responses. Biostatistics (in press)

19. Wilson loop expectations for finite gauge groups

[Student Paper Competition 1, (page 15)]

Sky CAO, *Stanford Statistics,Stanfod University*

Wilson loop variables are certain random variables which arise in physics. Understanding the expectations of these random variables is of particular interest. Recently, Chatterjee computed Wilson loop expectations to leading order for Ising lattice gauge theory at weak coupling. In this talk, I will introduce a generalization of this result to lattice gauge theories with finite gauge groups. In particular, the gauge group may be non-Abelian.

20 . Stochastic Generators with Global Spatio-Temporal Locally Diffusive SPDE Models

[IS 6, (page 5)]

Stefano CASTRUCCIO, *Stefano Castruccio,University of Notre Dame*
Geir-Arne FUGLSTAD, *Norwegian University of Science and Technology*

We propose a new class of models for globally resolved data as a solution of a locally diffusive SPDE model on the sphere and in time. By allowing the differential operator to change abruptly across large geographical descriptors such as land and ocean and smoothly across latitude and longitude, the model is able to capture the nonstationary behavior of many variables whose physics is dictated by both mesoscale and microscale processes, and is therefore used as a stochastic generator to assess the uncertainty of future climate projections. Furthermore, we show how inference can be performed effectively with millions of data points as the solution of the SPDE with finite elements can be expressed as a Gaussian Markov Random Field, thereby allowing sparsity in the linear algebra operations.

21. A Bayesian framework for sparse estimation in high-dimensional mixed frequency Vector Autoregressive models

[Student Paper Competition 1, (page 15)]

Nilanjana CHAKRABORTY, *Statistics,University of Florida*
Kshitij KHARE, *Department of Statistics, University of Florida*
George MICHAILIDIS, *Department of Statistics and the Informatics Institute, University of Florida*

This talk considers a Gaussian Vector Autoregressive model for high-dimensional mixed frequency data, where selective time series are collected at different frequencies. The high frequency ones are expanded and modelled as multiple time series, so as to match the low frequency sampling of the low frequency series. This leads to an expansion of the parameter space, thus posing challenges for estimation and inference in settings with limited number of observations. We address them by considering specific structural relationships in the representation of the high frequency series, together with sparsity of the ensuing model parameters through the introduction of spike-and-Gaussian slab prior distributions. In contrast to existing observation-driven methods, the proposed Bayesian approach accommodates general sparsity patterns, and makes a data-driven choice of the sparsity pattern. Under certain regularity conditions, we establish consistency for the posterior distribution under high-dimensional scaling. Applications on synthetic and real data illustrate the efficacy of the resulting estimates and corresponding credible intervals.

22 . SCALABLE BAYESIAN VARIABLE SELECTION AND GROUPINGFOR BINARY AND MULTICLASS OUTCOME DATA

[IS 42, (page 19)]

Sounak CHAKRABORTY, *Statistics,University of Missouri-Columbia*

In this paper, we consider Bayesian analysis of binary and multiclass support vector machines with featureselection. We consider the fully supervised support vector machine problem and place Gaussian spike andslab priors. We propose a computationally scalable Gibbs sampling algorithm, which has linear computationalcomplexity. We also consider Bayesian semi-supervised learning and propose a novel Bayesian approach forvariable selection with scalable Gibbs algorithm. Our proposed novel Gibbs sampler called Skinny Gibbswhich is much more scalable to high dimensional problems, both in memory and in computational efficiency.It can also avoid large matrix computations needed in standard Gibbs sampling algorithms. In terms of computational complexity for our Skinny Gibbs, it grows only linearly in the number of predictors. Efficiency ofour method for supervised and semi-supervised SVM models are demonstrated based on several simulationstudies and data analysis.

23. Model based methods to assess the impact of missing data on Oncology endpoints

[IS 61, (page 27)]

Arunava CHAKRAVARTTY, *Biostatistics/Novartis,Novartis*

Craig WANG, *Novartis*

n Oncology trials with survival endpoints large imbalances in the dropout rates can lead to a biased interpretation of study results. Such dropouts happen when patients withdraw from the study, often due to underlying side effects of the drug or desire to switch over other cancer therapies. In such cases an often used sensitivity analysis is to consider such dropouts as events to assess their impact on the study results. However such imputations may be too extreme by assuming events would occur at the time of censoring, thus fail to do a realistic impact assessment.

In this talk we present a statistical framework to assess the robustness of the time to event endpoints in oncology clinical trials based on sensitivity analyses under different scenarios of the dropout mechanism as an alternative to the extreme case scenario. A tipping point framework is also proposed to explore the extreme boundary conditions of the dropout mechanism that could lead to the study results to loose statistical significance. We will demonstrate the method based on a case study from an Oncology trial, assessing the impact of such dropouts on the the robustness of study results .

24 . Generating Actionable Insights from Real World Data during COVID-19 Pandemic

[Special Invited Session 4, (page 25)]

Aloka CHAKRAVARTY, *Immediate Office of the Commissioner, FDA,US Food and Drug Administration*

Clinical data collected outside of traditional clinical trialsalso known as Real-World Data (RWD) can provide insights to FDA on how COVID-19 treatments, diagnostics, and vaccines are performing in a variety of settings. The evidence generated depend on rigorous analytical methods as well as validation and cross-checking of analyses. During the pandemic, Evidence Accelerator brought together leading experts in health data aggregation and analytics in a unified, collaborative effort to share insights, compare results and answer key questions to inform the collective COVID-19 response. It provided a collaborative space for key players (nearly 200 organizations, to-date) across the health data ecosystem: FDA, major health data/technology organizations, academia, professional societies, health systems, insurers, drug and device industries, to assimilate and evaluate data generated from across the country. Dedicated parallel analyses workstreams focused on therapeutics and diagnostics, data and analytics. Participants advanced methods to leverage RWD for actionable insights. Three initial research areas (the use of hydroxychloroquine and azithromycin in hospitalized patients; the use of remdesivir; and the natural history of coagulopathy in COVID-19 patients) enabled establishing methodologies and processes for creating common data elements and interoperability. The comparable effort in Diagnostics explored the use of molecular and antibody tests. A critical early result of the Evidence Accelerator has been the characterization of the natural clinical history of COVID-19 in hospitalized patientsfoundational to ensuring testing performance, identifying treatment, predicting immunity, detecting potential for future waves of infection, and tracking mutation. In this presentation, some regulatory experience in this area will be discussed, along with case examples. In addition, some lessons learnt on best practices going forward will also be shared.

25 . Blocked adaptive randomization: how to do RAR if you really must.

[IS 55, (page 25)]

Richard CHAPPELL, *U Wisconsin Dept. of Biostatistics and Medical Informatics,University of Wisconsin*
Thevaasinen CHANDERENG, *Columbia University*

Response adaptive randomization (RAR) is the strategy of allowing subjects' randomization ratios in a clinical trial to depend on the outcomes of previous patients. It was originally proposed by Zelen (1979) as the "Randomized Play the Winner Rule" and subsequently criticized by Peto (1985) as unnecessarily prolonging a trial's conclusions and biasing results. The debate has not subsided since. This talk will give some background for RAR, describe the nature of the bias as resulting from confounding with time, and present some simple solutions.

Its intended audience is anyone who is interested in randomized clinical trials.

26. On High Dimensional, Robust, Unsupervised Record Linkage

[IS 34, (page 16)]

Ansu CHATTERJEE, *School of Statistics, University of Minnesota*

We develop a technique for record linkage on high dimensional data, where the two datasets may not have any common variable, and there may be no training set available. Our methodology is based on sparse, high dimensional principal components. Since large and high dimensional datasets are often prone to outliers and aberrant observations, we propose a technique for estimating robust, high dimensional principal components. We present theoretical results validating the robust, high dimensional principal component estimation steps, and justifying their use for record linkage. Some numeric results and remarks are also presented.

27. REVIEW ON MODEL DIAGNOSTICS FOR INCOMPLETE DATA

[IS 27, (page 13)]

Arkendu CHATTERJEE, *Global Biometrics and Data Science, BMS,Associate Director, BMS*

I will review the model diagnostics for incomplete longitudinal data. I will describe the local influence method introduced by (Verbeke, 2001) and explain its relation with index of local sensitivity to non-ignorability method (troxel, 2004). I will explain the graphical approaches introduced by Dobson (2003) and will make connections between their work and the local influence approach. Finally will review the model assessment based on observed replications.

28 . Individualized Risk Prediction: Lessons from Genetics and COVID-19

[Plenary Lecture 1, (page 3)]

Nilanjan CHATTERJEE, *615 N WOLFE ST, Suite E3527,Johns Hopkins University*
Jin JIN, *Johns Hopkins University*
Prosenjit KUNDU, *Johns Hopkins University*
Neha AGARWALA, *University of Maryland*
Haoyu ZHANG, *Harvard University*

Risk prediction models are central to the development of strategies for precision medicine, where risks and benefit need to be weighed at an individual level to implement medical interventions. In this talk, I will provide some broad perspectives on risk prediction from two contemporary applications (1) polygenic risk prediction and (2) COVID-19. I will describe a common theme regarding how comprehensive model development requires integration of information from multiple disparate studies. I will describe recent methodological development in statistical genetics as well as in some other fields that allows comprehensive model building using summary-level data, i.e. using estimates of parameters associated with a series of "reduced" models. I will then further address another common theme on how the evaluation of clinical utility of models requires going much beyond commonly used measures such as the area under the curve (AUC). Finally, I would share some perspectives regarding careful considerations needed before using statistical models to recommend health policy decisions.

29. Piecewise Polynomial Estimation on Grids by Dyadic CART and Optimal Regression Tree

[IS 17, (page 9)]

Sabyasachi CHATTERJEE, *Statistics, University of Illinois at Urbana-Champaign*
Subhajit GOSWAMI, *Tata Institute of Fundamental Research*

Proposed by Donoho (1997), Dyadic CART is a nonparametric regression method which computes a globally optimal dyadic decision tree and fits piecewise constant functions in two dimensions. In this article we define and study Dyadic CART and a closely related estimator, namely Optimal Regression Tree (ORT), in the context of estimating piecewise smooth functions in general dimensions in the fixed design setup. More precisely, these optimal decision tree estimators fit piecewise polynomials of any given degree. Like Dyadic CART in two dimensions, we reason that these estimators can also be computed in polynomial time in the sample size N via dynamic programming. We prove oracle inequalities for the finite sample risk of Dyadic CART and ORT which imply tight risk bounds for several function classes of interest. Firstly, they imply that the finite sample risk of ORT of order $r \geq 0$ is always bounded by Ck log N N whenever the regression function is piecewise polynomial of degree r on some reasonably regular axis aligned rectangular partition of the domain with at most k rectangles. Beyond the univariate case, such guarantees are scarcely available in the literature for computationally efficient estimators. Secondly, our oracle inequalities uncover minimax rate optimality and adaptivity of the Dyadic CART estimator for function spaces with bounded variation. We consider two function spaces of recent interest where multivariate total variation denoising and univariate trend filtering are the state of the art methods. We show that Dyadic CART enjoys certain advantages over these estimators while still maintaining all their known guarantees.

30. Integrating GWAS and multi-omics QTL summary statistics to elucidate disease genetic mechanisms via a hierarchical low-rank model

[IS 48, (page 22)]

Lin CHEN, *Department of Public Health Sciences, University of Chicago*
Yihao LU, *University of Chicago*
Jin LIU, *University of Chicago*

In the post-GWAS era, the functional mechanisms of trait-associated SNPs were extensively studied and evidences suggested that many of them may affect complex traits/diseases through their effects on expression levels and other omics traits such as DNA methylation levels. Extensive evaluations of genetic effects on omics traits have revealed an abundance of quantitative trait loci for omics traits (omics QTLs). QTL effects were reported to often have a bimodal tissue-sharing pattern – many common eQTLs having effects shared across tissues/conditions while many other QTLs having effects specific to certain (and potentially disease-relevant) tissue/cell types. With the availability of rich resources on GWAS and omics QTL summary statistics from different omics data types and different tissue types, in this work we propose an integrative methods for jointly analyzing GWAS and multiple sets of omics QTL summary statistics accounting for the hierarchical structure underlying omics QTLs. Here we propose an integrative analysis method that model the hierarchical low-rank structure of the latent association status between SNPs and tissue types for various omics data types. The proposed method was motivated by and was applied to analyses of multi-tissue eQTL and methylation QTL statistics from the Genotype-Tissue Expression (V8) project.

31. Digital endpoint development in atopic dermatitis clinical trials: using deep learning to quantify nocturnal scratch based on actigraphy data

[IS 20, (page 10)]

Xiaotian CHEN, *Data and Statistical Science, AbbVie*
Li WANG, *Data and Statistical Science, AbbVie*

In recent decades, novel digital endpoints emerged in clinical trial development. They complement traditional endpoints by inducing more objective measures and describing clinically meaningful concepts while maintaining consistency with traditional endpoints. In this presentation we will review the landscape of the digital endpoint development in atopic dermatitis therapeutic area and the state-of-art analytical methodologies on quantifying and predicting nocturnal scratch. As a huge amount of actigraphy data are often collected by wrist accelerometer digital devices in such trials, there has been increasing interest in utilizing deep learning techniques for accurate scratching detection based on actigraphy data. We focus on the application of recurrent neural network (RNN) that are widely used for the prediction problem in such situations to deal with long-range temporal dependencies in time series data. We also discuss the strategy in the conduct of atopic dermatitis clinical trials that reliably develop and validate nocturnal scratch algorithm.

32 . Mixed membership stochastic blockmodels for heterogeneous networks

[IS 36, (page 17)]

Yuguo CHEN, *Department of Statistics,University of Illinois at Urbana-Champaign*

Heterogeneous networks are useful for modeling complex systems that consist of different types of objects. We formulate a heterogeneous version of the mixed membership stochastic blockmodel to accommodate heterogeneity in the data and the content dependent property of the pairwise relationship. We also apply a variational algorithm for posterior inference. The proposed procedure is shown to be consistent for community detection under mixed membership stochastic blockmodels for heterogeneous networks. We demonstrate the advantage of the proposed method in modeling overlapping communities and multiple memberships through simulation studies and applications to a real data set.

33. Bayesian Estimation and Comparison of Conditional Moment Models

[Special Invited Session 1, (page 7)]

Siddhartha CHIB, *Olin Business School,Washington University in Saint Louis*
Minchul SHIN, *Federal Reserve Bank, Philadelphia*
Anna SIMONI, *CREST, CNRS, Ecole Polytechnique, Paris*

We consider the Bayesian analysis of models in which the unknown distribution of the outcomes is specified up to a set of conditional moment restrictions. The nonparametric exponentially tilted empirical likelihood (ETEL) function is constructed to satisfy a sequence of unconditional moments based on an increasing (in sample size) vector of approximating functions (such as tensor splines based on the splines of each conditioning variable). For any given sample size, results are robust to the number of such expanded moments. The posterior distribution is shown to satisfy the Bernstein-von Mises theorem, subject to a growth rate condition on the number of approximating functions, even under misspecification of the conditional moments. A large-sample theory for comparing different conditional moment models is developed. The central result is that the marginal likelihood criterion selects the model that is less misspecified. We also introduce sparsity-based model search for high- dimensional conditioning variables, and provide efficient MCMC computations for high-dimensional parameters. Along with clarifying examples, the framework is illustrated with real-data applications to risk-factor determination in finance, and causal inference under conditional ignorability.

34 . Randomization-only Inference in Experiments with Interference

[IS 54, (page 24)]

David CHOI, *Heinz College,Carnegie Mellon University*

In experiments that study social phenomena, such as peer influence or herd immunity, the treatment of one unit may influence the outcomes of other units. Such “interference between units” violates traditional approaches for causal inference, so that additional assumptions are required to model the underlying social mechanism. We propose a novel approach that requires no such assumptions, and hence may be useful in settings where interference is poorly understood. Our approach estimates differences in attributable effects (such as the difference between receiving the treatment directly, or receiving the control but having treated peers), with generalization to matching, weighting, and regression-based comparisons.

35 . Learning directed acyclic graphs based on single cell RNAseq data

[IS 18, (page 10)]

SHRABANTI CHOWDHURY, *Genetics and Genomic Sciences,Icahn school of Medicine at Mount Sinai*

Inferring gene/protein regulatory network based on multi-omics profiles is commonly pursued in biomedical studies. Specifically, directed acyclic graph (DAG) constructed based on -omics data has recently been used to infer causal associations among the interacting genes. Single cell RNA-seq data provides unique opportunities to characterize common and differential molecular networks across different cell types. However, unlike bulk RNA-seq profiling, many transcripts in single-cell RNA-seq experiments go undetected, leading to sparsity, which then imposes key challenges in effective network analysis. To address this issue, we develop a new method scDAGBagM to construct DAGs based on scRNAseq data. Specifically, in scDAGBagM, for each gene, we (1) use a pair of binary and continuous nodes to represent its expression status and level respectively and (2) make the binary node a natural parent of the corresponding continuous node. In this way, along with characterizing the zero-inflated distribution of the data, we can achieve better power to detect the gene-gene regulations that are largely signaled through active/inactive statuses of genes. We then utilize a modified hill-climbing algorithm to build DAGs. scDAGBagM employs a novel aggregation procedure inspired by bootstrap aggregating to tackle the high variability in DAG structure learning even with moderate number of nodes. scDAGBagM is also flexible in taking into prior information of edge directions which can be important for DAG structure learning as edge directions are not always identifiable without external information. We illustrate the performance of scDAGBagM through simulation studies and a real scRNAseq data set of a lung cancer study.

36. Bayesian Hierarchical Spatial Models for Small Area Estimation

[IS 13, (page 8)]

Hee Cheol CHUNG, *Department of Statistics,Texas A&M University*
Gauri Sankar DATTA, *University of Georgia*

For over forty years, the Fay-Herriot model has been extensively used by National Statistical Offices around the world to produce reliable small area statistics. This model develops prediction of small area means of a continuous outcome of interest based on a linear regression on suitable auxiliary variables. Model errors, also known as small area effects, of the Fay-Herriot model are treated as independent and normally distributed zero-mean random variables with an unknown variance. Often population means of geographically contiguous small areas display a spatial pattern. The independence assumption for the random effects may not hold when effective auxiliary variables are unavailable. Lack of suitable covariates to account for the variation of the geographic domain means results in a spatial pattern among the random effects. We consider several spatial random-effects models, including the popular conditional autoregressive (CAR) and simultaneous autoregressive (SAR) models as alternatives to the Fay-Herriot model. We carry out a Bayesian analysis of these models based on a class of popular noninformative improper prior densities for the model parameters. We assess the effectiveness of these spatial models based on a simulation study and a real application. We consider the prediction of statewide four-person family median incomes for the U.S. states based on the 1990 Current Population Survey and the 1980 Census. We assess the accuracy of our predictions against the corresponding 1990 Census values. In some applications, small areas are formed after surveys, and these areas have no sample data. Proposed spatial models generate better predictions of unsampled small area means by borrowing from neighboring residuals than the synthetic regression means from regular independent random effects model. For all the spatial models considered, their posterior distributions based on a useful class of improper prior densities on model parameters, even in the absence of data from some small areas, are shown to be proper under the same set of mild conditions.

37. Simulation Practices for Adaptive Clinical Trial Design in Drug and Device Development

[IS 2, (page 3)]

Greg CICCONETTI, *Teri Anderson,AbbVie*

Following the commitments of the PDUFA VI initiative, the FDA recently issued a revised guidance on adaptive designs, prominently emphasizing the role played by simulation in designing complex clinical trials. Simulation is a well-recognized tool in designing adaptive trials, as evidenced by number and variety of publications on the subject. However, there is still lack of clarity and consistency regarding what constitutes a good back bone of a simulation process and how to document it. The latter is of crucial importance in situations when the simulation report becomes a key design justification document. To address this need, a group of industry statisticians with extensive experience in designing adaptive trials got together (under the sponsorship of DIA Adaptive Design Working Group) and summarized a core set of requirements constituting good simulation practices for a few types of commonly used adaptive designs. While the Working Groups output was motivated by the goal of creating a quality simulation report, this presentation will focus on the proper planning and customization of the simulation experiment to the trial design or problem at hand. It will discuss key components of trial simulation, and link questions of interest to statistical models and assumptions and appropriately documenting them in order to facilitate discussion within cross-functional teams.

38. Application of polygenic risk scores to diverse populations

[IS 65, (page 29)]

Brandon COOMBES, *Department of Quantitative Health Sciences, Mayo Clinic*
Anthony BATZER, *Mayo Clinic*
Gregory JENKINS, *Mayo Clinic*
Euijung RYU, *Mayo Clinic*

Polygenic risk scores (PRS), a machine learning approach in genetics, are increasingly being used in research to predict disease and show polygenic overlap between traits. However, PRS trained in one ancestry and applied to another ancestry can suffer reduced accuracy due to differences in minor allele frequency and linkage disequilibrium across the genome. A recent study found that almost 80% of participants included in genome-wide association studies (GWAS) to date are of European ancestry. Thus, for most traits, GWAS performed in diverse ancestries are either non-existent or have much smaller sample sizes and thus worse predictive power for PRS. While a few methods have been developed to improve the accuracy of cross-ancestry PRS, this remains an open research topic. Here, we explore multiple ways to estimate the PRS for type II diabetes (T2D) in the Sangre por Salud (SPS) Biobank, a large Hispanic biobank (N 4000), to predict a variety of biomarkers including hemoglobin A1c, a biomarker that is elevated among individuals with T2D. In this Hispanic sample, we compare the performance of traditional PRS derived from either a European (N = 1.11M), Hispanic (N = 20K), African American (N = 56K), Asian (N = 216K) or trans-ethnic (N = 1.4M) GWAS of T2D. We then compare this performance to cross-ancestry PRS methods (PolyPred and PRS-CSx) which leverage the learned genetic architecture of the smaller Hispanic GWAS to improve prediction performance. While these cross-ancestry PRS methods have shown that they can improve PRS prediction, they are much more computationally intensive and much harder to implement than traditional PRS. Thus, we propose a computationally fast and simple alternative to combine the traditional PRS across ancestries using a principal component approach.

39 . Latent Trait Shared Parameter Mixed Models for Missing Ordinal Ecological Momentary Assessment Data

[IS 60, (page 27)]

John CURSIO, *Public Health Sciences,University of Chicago*

Latent trait shared parameter mixed-models (LTSPMM) for missing Ecological Momentary Assessment (EMA) data are developed in which two mood outcomes, negative and positive affect are collected in an intermittent fashion. Using Item Response Theory (IRT) models, a latent trait is used to model the missingness mechanism and estimated jointly with a mixed-model for longitudinal ordinal outcomes. Both one- and two-parameter LTSPMM are presented. These LTSPMMs offer a novel way to analyze EMA data with many unique response patterns that are assumed to be missing not at random. Previously, these LTSPMM were used with normal outcomes, and here ordinal models are estimated and compared to missing at random mixed-models. Item intercept and discrimination parameters, latent traits, and mixed-model covariates will be shown using both model types. The models show that the latent trait of responsiveness is significantly related to both negative and positive affect and that subjects that were more responsive had better moods. Finally, a discussion of model estimation issues will be presented for the ordinal LTSPMM.

40 . Motif Estimation via Subgraph Sampling

[Student Paper Competition 1, (page 15)]

Sayan DAS, *Department of Mathematics,Columbia University*
Bhaswar B. BHATTACHARYA, *University of Pennsylvania*
Sumit MUKHERJEE, *Columbia University*

Network sampling is an indispensable tool for understanding features of large complex networks where it is practically impossible to search over the entire graph. In this talk, we discuss a framework for statistical inference for counting network motifs, such as edges, triangles, and wedges, in the widely used subgraph sampling model, where each vertex is sampled independently, and the subgraph induced by the sampled vertices is observed. We derive necessary and sufficient conditions for the consistency and the asymptotic normality of the natural Horvitz-Thompson (HT) estimator, which can be used for constructing confidence intervals and hypothesis testing for the motif counts based on the sampled graph. In particular, we show that the asymptotic normality of the HT estimator exhibits an interesting fourth-moment phenomenon, which asserts that the HT estimator (appropriately centered and rescaled) converges in distribution to the standard normal whenever its fourth-moment converges to 3 (the fourth-moment of the standard normal distribution).

41 . Pseudo-Bayes Small Area Estimation via Compromise Regression Weights

[IS 45, (page 21)]

Gauri Sankar DATTA, *Department of Statistics,University of Georgia/US Census Bureau*
Lee JUHYUNG, *University of Georgia*
Li JIACHENG, *University of Georgia*

Model-based estimate of a small area mean is obtained by shrinking a noisy direct estimate to a regression synthetic estimate based on a model. If a model is mis-specified, model-based estimates of areas with less reliable direct estimates may be suboptimal due to their overreliance on a poorly estimated model. Jiang et al. (2011, JASA) and Nicholas et al. (2020) proposed frequentist estimation of the model by minimizing an estimated total mean squared error (ETMSE). As an alternative to the solutions suggested by these authors, we pursue a Bayesian approach. We suitably standardize the ETMSE to construct a pseudo-likelihood for the model parameters and use a class of noninformative priors to derive Bayesian estimates of small area means. We apply the propose method to estimate median incomes of U.S. states. Application and a simulation study show that our Bayesian solution competes favorably with the frequentist methods when assessed based on suitable frequentist criteria.

42. An Accurate Coreset Methodology for Efficient Reduction of Spatial Data

[IS 45, (page 21)]

Ranadeep DAW, *University of Missouri, Department of Statistics,University of Missouri*
Christopher K WIKLE, *University of Missouri*

A coreset is an effective summary of the original partially redundant data such that the solution of a problem over the coreset yields results similar to that obtained with the original dataset. Although the general notion of a coreset is clear, there are many different approaches to select a coreset based on the type of problem being considered. A recent approach to formalize coreset construction is based on the notion of an accurate coreset, which selects a coreset in such a way that the solution of a statistical problem (central tendency measurement, linear model estimation, dimension reduction, loss minimization, etc.) over the accurate coreset is exactly same as the solution obtained from the full set. Here, we propose such a coreset construction for the spatial data under a Bayesian formulation. We demonstrate the methodology on point (environmental) and areal (federal survey) data and discuss the similarity to other recent approaches for optimal reduction of spatial data.

43. Connectivity Regression

[Student Paper Competition 2, (page 19)]

Neel M. DESAI, *Rice University*
Veera BALADANDAYUTHAPANI,
Jeffrey MORRIS,

Assessing heterogeneity of multivariate associations across covariates is an important problem in many areas of modern science, including in neuroscience to discover factors explaining inter-subject variability in functional connectivity networks. In this work, we present general methodology to regress subject-specific networks on a set of covariates that produces multiplicity-adjusted hypothesis tests for which covariates affect the networks, as well as statistical measures indicating which network edges are driving these differences. Our strategy involves projecting a subject-specific empirical correlation matrix into the Fisher correlation space using a matrix logarithm transform, which ensures positive-semi definiteness and justifies Gaussian modeling. Using a Gaussian multivariate regression framework in this space with cutting-edge sparsity priors, we regress the networks on predictors while discovering and accounting for second-order dependence across network edges which we show leads to greater efficiency and power for statistical inference using the principles of Seemingly Unrelated Regression. We apply our approach to analyze functional connectivity networks of 1003 healthy young patients taken from the Human Connectome Project (HCP), finding subject-specific connectivity is associated with their post central gyrus area, precuneus area, and language processing capabilities. Overall, our framework serves as a promising tool for solving problems in functional connectivity and addresses a growing need for performing inference on observations with complex structure.

44. Bayesian Model Assessment and Selection Using Bregman Divergence

[IS 29, (page 14)]

Dipak DEY, *Department of Statistics,University of Connecticut*
Gyuhyeong GOH, *Kansas State University*

One of the fundamental steps in statistical modeling is to select the best-fitting model from a set of candidate models for given data. In this presentation, based on Bayesian decision theory, we introduce a new model selection criterion, called Bregman divergence criterion (BDC). The proposed criterion improves many existing Bayesian model selection methods such as Bayes factor, intrinsic Bayes factor, pseudo-Bayes factor, etc. In addition, using a Monte Carlo approach, we develop an efficient estimator that significantly eases the computational burden associated with our approach and prove its consistency. The versatility of our methodology is demonstrated on both simulated and real data; to this end, some illustrative examples are provided for linear regression models and longitudinal data models.

45. EM based approach for analysis of multi-platform genomics data

[IS 42, (page 19)]

Tanujit DEY, *Center for Surgery and Public Health, BWH,Harvard Medical School*
Sounak CHAKRABORTY, *Department of Statistics, University of Missouri*
Hao XUE, *Department of Biostatistics, Harvard School of Public Health*

Combining genomics data across different platforms with patient's clinical outcomes improves precision and accuracy in the identification of corresponding biomarkers. It helps us reveal the complex biological mechanisms behind the development of the disease. We propose an integrativeBayesian model for detecting genes that have significant associations with clinical outcomes. This framework integrates data from different platforms in a hierarchical manner.To evaluate the performance of our model, we apply it to analyze both simulated data and a real data set on cancer genomics.

46 . A matrix-free profile likelihood method for high-dimensional factor model

[IS 66, (page 29)]

Somak DUTTA, *Statistics,Iowa State University*
Fan DAI, *Michigan Technological University*
Ranjan MAITRA, *Iowa State University*

In this talk we discuss a profile likelihood method for estimating the covariance parameters in a Gaussian factor model with fewer observations than number of variables. We develop a fast matrix-free computation framework based on an implicitly restarted Lanczos algorithm and a limited memory quasi-Newton optimization algorithm. We demonstrate that our approach is substantially faster than the EM algorithm for high dimensional problems. We illustrate our method on a fMRI dataset on suicide attempters, suicide ideators and normal subjects. We conclude with some discussion on extending factor models to non-Gaussian data.

47. The two strategies in estimating the values of dynamic treatment regimes

[IS 14, (page 8)]

Yixin FANG, *Data and Statistical Sciences,AbbVie*

In precision medicine, it is crucial to estimate the value associated with a dynamic treatment regime. In this talk, we start with reviewing two most commonly used strategies in the literature of causal inference, the weighting strategy and the standardization strategy. Then we extend these two strategies to estimate the value of an individualized treatment regime and the value of a dynamic treatment regime, respectively. Although these two strategies are quite different at the beginning, at the end both of them are doubly robust methods for estimating the treatment effects under the same set of identifiability conditions.

48 . Considerations in using External Control from Real-world Data to Sup-

port FDA Approvals

[IS 30, (page 14)]

Dai FENG, *GMA Statistics, DSS, AbbVie,AbbVie*
Meijing WU, *AbbVie*
Hongwei WANG, *AbbVie*
Yixin FANG, *AbbVie*
Weili HE, *AbbVie*

With a long history of using real-world evidence (RWE) to monitor and evaluate the safety of postmarketing drug products, the FDA is committed to expanding the use of RWE for decisionmaking on drug efficacy. In 2020, 75% of FDA-approved New Drug Applications (NDAs) and Biologics License Applications (BLAs) included a RWE study, jumping from 49% in 2019. The focus of this talk is on using external controls from real-world data (RWD) in randomized clinical trials (RCTs) to establish efficacy of drugs. We will describe some lessons learned examples from recent NDA and BLA submissions, aiming at elucidating how the evidence provided can be considered as substantial. Furthermore, we will discuss practical considerations in using external control from RWD, which include the rationale and detailed steps for implementation in the trial design, execution, and analysis stages.

49 . Novel dynamic multiscale spatiotemporal models for multivariate Gaussian data with applications to stratospheric temperatures

[IS 51, (page 23)]

Marco FERREIRA, *Marco Ferreira,Department of Statistics, Virginia Tech*
Mohamed ELKHOULY, *Department of Statistics, University of Wisconsin - Madison*

We propose a novel class of multiscale spatiotemporal models for multivariate Gaussian data. First, we decompose the multivariate data and the underlying latent process with a novel generalized multiscale Haar decomposition. We then assume that the resulting latent multiscale coefficient matrices evolve through time with matrix-variate state-space equations. This flexible model framework allows for both stationary and nonstationary latent processes. Further, we develop a singular matrix-variate forward filter backward sampler for efficient posterior exploration. Importantly for practical purposes, our proposed multiscale spatiotemporal algorithm scales linearly with dataset size and is fully parallelizable. To illustrate the usefulness and flexibility of our dynamic multivariate multiscale framework, we present an application to a spatiotemporal NCEP/NCAR Reanalysis-I dataset on stratospheric temperatures over North America from 1951 to 2016.

50 . Conditional calibration for FDR control under dependence

[IS 16, (page 9)]

Will FITHIAN, *Statistics, UC Berkeley,UC Berkeley Statistics*
Lihua LEI, *Stanford University*

We introduce a new class of methods for finite-sample false discovery rate (FDR) control in multiple testing problems with dependent test statistics where the dependence is fully or partially known. Our approach separately calibrates a data-dependent p-value rejection threshold for each hypothesis, relaxing or tightening the threshold as appropriate to target exact FDR control. In addition to our general framework we propose a concrete algorithm, the dependence-adjusted Benjamini-Hochberg (dBH) procedure, which adaptively thresholds the q-value for each hypothesis. Under positive regression dependence the dBH procedure uniformly dominates the standard BH procedure, and in general it uniformly dominates the BenjaminiYekutieli (BY) procedure (also known as BH with log correction). Simulations and real data examples illustrate power gains over competing approaches to FDR control under dependence. This is joint work with Lihua Lei.

51 . Examples and Stories: Computer Scientists and Statisticians - What we need to learn from each other

[IS 14, (page 8)]

Haoda FU, *Eli Lilly and Company*

For the past 3 years, I have been the Enterprise Lead for Machine Learning and AI for Eli Lilly and Company. In this data scientist team, people from different discipline including computer science and statistics. I am going to share our journey and lessons learned on how these two team worked together and how we can learn from each other for better collaborations.

52 . Statistical Construct of Extrapolation: Composite Likelihood and Bayesian Approaches

[IS 32, (page 16)]

Margaret GAMALO, *GBDM-Inflammation and Immunology,Pfizer*

Depending on the degree of extrapolation, streamlining the pediatric drug development program is done with the awareness of the indication being pursued, the mechanism of action of the drug, and the strategic goal within a landscape to ensure timely access to drugs. With these considerations, there are a common development archetypes, e.g., investigational drug is (i) First-in-class or first-in-indication in adults, (ii) Established class where adult trials have confirmed safety and efficacy, achieving registration, (iii) Established class where adult and pediatric trials have been conducted, achieving registration. These archetypes imply different degrees of extrapolation and can be captured through an extrapolation coefficient, a quantitative elicitation of similarity of diseases that is independent from mere numerical similarity of treatment responses between cohorts, or by design. This presentation will focus on different aggregative methods to allow for extrapolation within the archetypes either through Bayesian hierarchical models or the use of composite likelihoods for the analysis of treatment effects across cohorts within the overall pediatric clinical development. The whole framework is balanced within the context of providing safe and effective medicines for children.

53 . Clustered-Temporal Bayesian Model for Brain connectivity in Neuroimaging data

[IS 52, (page 23)]

Nairita GHOSAL, *Merck & Co., INC.*
Sanjib BASU, *University of Illinois at Chicago*

Functional brain connectivity refers to temporal dependence of activation pattern of brain regions. Functional magnetic resonance imaging (fMRI) measures functional connectivity by observing co-activation pattern of anatomically separated brain regions in resting state. The time-dependent model for functional connectivity jointly considers the time sequence of fMRI measurement and the modularity structure of brain regions. We have applied dynamic linear model to capture the temporal structure of data and the potential correlation between connected regions are modeled using Hidden Potts model with latent variable. As an application we used these models to analyze functional connectivity in the Autism Brain Imaging Data Exchange (ABIDE) data set. Keywords: Autism, Dynamic Linear Model, Hidden Potts Model

54. Bayesian Modeling of North Atlantic Tropical Cyclone Activity

[IS 35, (page 17)]

Joyee GHOSH, *Statistics and Actuarial Science,The University of Iowa*
Xun LI, *The University of Iowa*
Gabriele VILLARINI, *The University of Iowa*

Seasonal forecasting of the frequency of North Atlantic tropical storms is of interest, because it can provide basic information towards improved preparation against these storms. It has been shown that sea surface temperatures during the hurricane season can predict tropical cyclone activity well. But predictions need to be made before the beginning of the hurricane season, when the predictors are not yet observed. Several climate models issue forecasts of the SSTs, which can be used instead. Such models use the forecasts of SSTs as surrogates for the true SSTs. In this work, we develop a fully Bayesian negative binomial regression model, which makes a distinction between the true SSTs and their forecasts, both of which are included in the model. For prediction, the true SSTs may be regarded as unobserved predictors and sampled from their posterior predictive distribution. Our model can simultaneously handle missing predictors and variable selection uncertainty. If the main goal is prediction, an interesting question is: should we include predictors in the model that are unavailable at the time of prediction? We attempt to answer this question using simulation studies. Based on the North Atlantic tropical storms dataset, we demonstrate that our model can provide gains in prediction, especially in the months closer to the hurricane season.

55. Adaptive Multi-Arm Multi-Stage Design

[IS 2, (page 4)]

Pranab GHOSH, *Biostatistics,Pfizer Inc.*

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

56. Possible Hazards of Proportional Hazards Models

[IS 44, (page 20)]

Sujit GHOSH, *Department of Statistics,North Carolina State University*

Alvin SHENG, *North Carolina State University*

The Cox proportional hazard (PH) model is widely used to determine the effects of risk factors and treatments on survival time of subjects that might be right censored. The selection of covariates depends crucially on the specific form of the conditional hazard model, which is often assumed to be PH, accelerated failure time (AFT), or proportional odds (PO). However, we show that none of these semiparametric models allow for the crossing of the survival functions and hence such strong assumptions may adversely affect the selection of variables. Moreover, the most commonly used PH assumption may also be violated when there is a delayed effect of the risk factors. Alternative models are presented which provides a smooth estimator of the conditional hazard that encompasses PH structure. Empirical results based on several simulated data scenarios indicate the superior performances of the proposed model and thereby shown to avoid possible hazards of the proportional hazard like assumptions.

57. Bootstrapping Lp-Statistics in High Dimensions

[IS 40, (page 18)]

Alexander GIESSING, *Department of Operations Research and Financial Engineering,Princeton University*

Jianqing FAN, *Princeton University*

In this talk we consider a new bootstrap procedure to estimate the distribution of high-dimensional Lp-statistics, i.e. the Lp-norms of the sum of n independent d-dimensional random vectors with $d >> n$ and $p \in [1, \infty]$. We provide a non-asymptotic characterization of the sampling distribution of Lp-statistics based on Gaussian approximation and show that the bootstrap procedure is consistent in the Kolmogorov-Smirnov distance under mild conditions on the covariance structure of the data. As an application of the general theory we propose a bootstrap hypothesis test for simultaneous inference on high-dimensional mean vectors. We establish its asymptotic correctness and consistency under high-dimensional alternatives, and discuss the power of the test as well as the size of associated confidence sets. We illustrate the bootstrap and testing procedure numerically on simulated data.

58. Multivariate spectral downscaling for PM2.5 species

[IS 6, (page 5)]

Yawen GUAN, *Statistics,University of Nebraska - Lincoln*

Brian J REICH, *North Carolina State University*

James A MULHOLLAND, *Georgia Tech*

Howard H CHANG, *Emory University*

Fine particulate matter (PM2.5) is a mixture of air pollutants that has adverse effects on human health. Understanding the health effects of PM2.5 mixture and its individual species has been a research priority over the past two decades. However, the limited availability of speciated PM2.5 measurements continues to be a major challenge in exposure assessment for conducting large-scale population-based epidemiology studies. The PM2.5 species have complex spatial-temporal and cross dependence structures that should be accounted for in estimating the spatiotemporal distribution of each component. Two major sources of air quality data are commonly used for deriving exposure estimates: point-level monitoring data and gridded numerical computer model simulation, such as the Community Multiscale Air Quality (CMAQ) model. We propose a statistical method to combine these two data sources for estimating speciated PM2.5 concentration. Our method models the complex relationships between monitoring measurements and the numerical model output at different spatial resolutions, and we model the spatial dependence and cross dependence among PM2.5 species. We apply the method to combine CMAQ model output with major PM2.5 species measurements in the contiguous United States in 2011.

59. Computing models with big data: privacy consideration, distributed implementation

[IS 6, (page 5)]

Rajarshi GUHANIYOGI, *Statistics, UC Santa Cruz*

Bayesian computation with large sample size and large number of variables present enormous challenges in the light of recently emerging applications in neuroimaging, environmental science and forestry, among others. Preserving privacy of data samples is another serious consideration while computing Bayes methods on medical record data containing sensible private information. In this talk, I will discuss two broad approaches we pursued recently. The first approach is based on data compression to simultaneously offer computational efficiency of Bayes models after preserving data privacy. The second approach develops a divide and conquer algorithm for scalable computation of big and correlated data. Theoretical results justifying these approaches will be discussed. Part of this work has been jointly conducted with Sanvesh Srivastava, Cheng Li, Terrance Savitsky, Laura Baracaldo and Sudipto Banerjee.

60. Convex Regression in High Dimensions

[IS 5, (page 4)]

Adityanand GUNTUBOYINA, *Statistics,University of California Berkeley*

Gil KUR, *Massachusetts Institute of Technology*

Fuchang GAO, *University of Idaho*

Bodhisattva SEN, *Columbia University*

The least squares estimator (LSE) is shown to be suboptimal in squared error loss in the usual nonparametric regression model with Gaussian errors for $d \geq 5$ for each of the following families of functions: (i) convex functions supported on a polytope (in fixed design), (ii) bounded convex functions supported on a polytope (in random design), and (iii) convex Lipschitz functions supported on any convex domain (in random design). For each of these families, the risk of the LSE is proved to be of the order $n^{2/d}$ (up to logarithmic factors) while the minimax risk is $n^{4/(d+4)}$, for $d \geq 5$. In addition, the first rate of convergence results (worst case and adaptive) for the full convex LSE are established for polytopal domains for all $d \geq 1$. Some new metric entropy results for convex functions are also proved which are of independent interest.

61. A Modified Graphical Approach with Generalized Sequentially Rejective Principle to Control Familywise Error Rate

[IS 43, (page 20)]

Wenge GUO, *Mathematical Sciences,New Jersey Institute of Technology*

Li YU, *Merck & Co.*

Various sequentially rejective, weighted Bonferroni-based multiple testing procedures (MTPs) have been applied in clinical trials for addressing multiple research questions. However, such MTPs may become complex with the increasing sources of multiplicity. In order to make the testing strategy more clear and intuitive to communicate with non-statisticians, graphical approaches have been proposed to visualize multiple test procedures. In this talk, we will firstly introduce a modified graphical approach which is more flexible and efficient than the existing graphical approaches. It allows one to reject more than one hypothesis at each step. We then generalize the sequential rejection principle by Goeman and Solari (2010) and prove the control of the familywise error rate under arbitrary dependence for the modified graphical approach. Through two examples, we finally illustrate the performance of the proposed graphical approach.

62. Confident predictions even when distributions shift

[Student Paper Competition 1, (page 15)]

Suyash GUPTA, *Statistics,Ph.D. student, Statistics, Stanford University*

Maxime CAUCHOIS, *Ph.D. student, Statistics, Stanford University*

Alnur ALI, *Post doctoral candidate, Electrical Engineering, Stanford University*

John DUCHI, *Assistant Professor, Statistics and Electrical Engineering, Stanford University*

While the traditional viewpoint in machine learning and statistics assumes training and testing samples come from the same population, practice belies this fiction. One strategycoming from robust statistics and optimizationis thus to build a model robust to distributional perturbations. In this talk, we take a different approach to describe procedures for robust predictive inference, where a model provides uncertainty estimates on its predictions rather than point predictions. We present a method that produces prediction sets (almost exactly) giving the right coverage level for any test distribution in an f-divergence ball around the training population. The method, based on conformal inference, achieves (nearly) valid coverage in finite samples, under only the condition that the training data be exchangeable. An essential component of our methodology is to estimate the amount of expected future data shift and build robustness to it; we develop estimators and prove their consistency for protection and validity of uncertainty estimates under shifts. By experimenting on several large-scale benchmark datasets, including Recht et al.s CIFAR-v4 and ImageNet-V2 datasets, we provide complementary empirical results that highlight the importance of robust predictive validity.

63. Aggregate Safety Analysis and Planning in Clinical Development

[IS 26, (page 13)]

Barbara HENDRICKSON, *Pharmacovigilance and Patient Safety, AbbVie,AbbVie*

The Drug Information Association - American Statistical Association sponsored Interdisciplinary Safety Evaluation working group has developed an Aggregate Safety Assessment Plan (ASAP) process. The ASAP describes an approach for product level safety planning and evaluation across the product life-cycle. This approach leverages dynamic multidisciplinary collaboration among statisticians, clinicians, epidemiologists, and other subject matter experts. Proactive and systematic aggregate safety assessment is essential for understanding a products evolving safety profile and for characterizing important product risks earlier in development. The ASAP promotes identification of the safety topics of interest for the product as well as definition of the epidemiology of the patient population and background rates of anticipated adverse events. Development of the ASAP prompts consideration of thoughtful standardized clinical trial data collection and analyses and key safety knowledge gaps. The presentation will discuss components of the ASAP template and how the ASAP enhances the ability to answer the fundamental safety questions for the product at filing.

64. Improved uncertainty quantification for random forests and other ensembles

[IS 59, (page 27)]

Giles HOOKER, *Statistics and Data Science,Cornel University*
Zhengze ZHOU, *Cornell University*
Indrayudh GHOSAL, *Cornell University*

This talk discusses uncertainty quantification and inference using ensemble methods. Recent theoretical developments inspired by random forests have cast bagging-type methods as U-statistics when bootstrap samples are replaced by subsamples, resulting in a central limit theorem and hence the potential for inference. However, to carry this out requires estimating a variance for which all proposed estimators exhibit substantial upward bias. In this talk, we convert subsamples without replacement to subsamples with replacement resulting in V-statistics for which we prove a novel central limit theorem. We also show that in this context, the asymptotic variance can be expressed as the variance of a conditional expectation which is approximated by sampling from the empirical distribution and allows for valid bias corrections. We finish by illustrating the use of these tools in combining or comparing statistical models.

65. The Full Picture: Making the Data Insight a Reality

[IS 26, (page 13)]

Erya HUANG, *Clinical Statistics,Bayer U.S. LLC*

Clinical trial databases contain abundant information with complex interrelations, increasing the demand for innovative data visualization methods that provide an insightful overview of study results. In this presentation, we will provide an overview of eight data visualization tools developed by the Biostatistics Innovation Center at Bayer AG. Demo of one tool, Subgroup Explorer, will be provided as an example. Subgroup Explorer is an application to do hundred thousand of subgroup analyses in one go: identifying outcome relevant subgroups is made as simple as possible. This reassures the clinical team that no potential subgroup effect has been overlooked or leads to targeted planning of future trials. R package for most tools, including subscreen are available on CRAN.

66. Causal Machine Learning for predictive biomarker identification

[IS 14, (page 8)]

Xin HUANG, *Discovery and Exploratory Statistics (DIVES), Data & Statistical Sciences,AbbVie Inc.*

Biomarkers are the foundation of precision medicine. The identification of prognostic and predictive biomarkers is an important scientific component in advancing the drug discovery and development pipeline. Many machine learning methods have been developed to identify important prognostic biomarkers. However, most existing algorithms are not directly applicable for identifying predictive biomarkers because individual treatment effect is not observable. In this presentation, we focus on the discussion of how to modify popular ensemble learning methods and use off-the-shelf machine learning software to identify important predictive biomarkers for continuous, binary, and time-to-event endpoints. Simulation studies are performed to compare different methods. Finally, a real example is used to demonstrate on how to use the proposed methods for successful subgroup identification for an immunological disease treatment.

67 . Survival Trees for Informative Interval-Censored Data

[IS 24, (page 12)]

Noorie HYUN, *Division of Biostatistics,Medical College of Wisconsin*

Xiao LI, *Medical College of Wisconsin*

For a disease screening cohort, accurately modeling individual time-to-event can help determine the appropriate screening or treatment triage. The clinic visit interval for screening a disease is generally broad, so the exact disease onset time is unobserved. Instead, the time interval between the latest normal and earliest abnormal disease status is observed. Moreover, the visiting process can for screening a disease can be correlated with the time-to-event because subjects at higher risk have faster follow-up than subjects at lower risk. While resolving these issues, we are motivated to develop a survival tree directly predicting the time-to-event rather than the probability to directly incorporate the final leaves information into determining the next follow-up time. We model the visiting process and include the model-based prediction as a latent variable in a primary accelerated failure time model for time-to-disease. We use nonparametric partial rank score test statistic for variable selections and deviance from the restricted mean survival time in each leaf for splitting nodes under the accelerated failure time model framework. The developed survival tree algorithm is evaluated via a simulation study.

68. Bayesian Interim Decision Strategy in Phase III Time-To-Event Study

[IS 20, (page 10)]

David IPE, *Data and Statistical Science, AbbVie*

Saurabh MUKHOPADHYAY, *Data and Statistical Science, AbbVie*

Accumulating data in long-term clinical trials can be very informative in decision making, particularly if there is a short-term surrogate endpoint. However, there are many practical and technical challenges to make such assessments. We propose a unified Bayesian decision-making framework to assess the accumulating evidence in an ongoing trial. Given the accumulated data, the posterior probability of treatment effect or the predictive probability of study success can be useful for capturing the information in a unified manner. Using a virtual trial framework with time-to-event endpoint, we show how to determine in advance what would be a good time for such interim decision-making assessment (IDA) and whether a second assessment would be needed. We also illustrate how to determine optimal thresholds for this decision-making. We then assess the operating characteristics of the entire process

69. High-dimensional quadratic classifiers under the strongly spiked eigenvalue model

[IS 28, (page 14)]

Aki ISHII, *Department of Information Sciences,Tokyo University of Science*

Kazuyoshi YATA, *Institute of Mathematics, University of Tsukuba*

Makoto AOSHIMA, *Institute of Mathematics, University of Tsukuba*

One of the features of modern data is that the data dimension is extremely high, however, the sample size is relatively low. As for the high-dimensional data, it is very important to construct theories on the basis of the eigenstructures. There are two types of high-dimensional eigenvalue models: the strongly spiked eigenvalue (SSE) model and the non-SSE (NSSE) model. In this talk, we consider high-dimensional quadratic classification under the SSE model. We give new classifiers by using a data transformation technique. We show that our classifiers have preferable properties in theory. Finally, we demonstrate our classifiers by using microarray data sets.

70. Optimal Crossover Designs for Generalized Linear Models

[IS 56, (page 26)]

Jeevan JANKAR, *Department of Statistics,University*

of Georgia, Athens
Abhyuday MANDAL, *University of Georgia, Athens*
Jie YANG, *University of Illinois, Chicago*

We identify locally D-optimal crossover designs for generalized linear models. We use generalized estimating equations to estimate the model parameters along with their variances. To capture the dependency among the observations coming from the same subject, we propose six different correlation structures. We identify the optimal allocations of units for different sequences of treatments. For two-treatment crossover designs, we show via simulations that the optimal allocations are reasonably robust to different choices of the correlation structures. We discuss a real example of multiple treatment crossover experiments using Latin square designs. Using a simulation study, we show that a two-stage design with our locally D-optimal design at the second stage is more efficient than the uniform design, especially when the responses from the same subject are correlated.

71. Efficient Trial Design Selection in Rare Disease Clinical Development

[IS 19, (page 10)]

Yannis JEMIAI, *Cytel,Cytel*

Designing a clinical trial is a complex endeavor that involves many assumptions about unknown factors. Ensuring robustness of the trial design to its real world implementation requires simulations at scale to assess sensitivity to assumptions and impact on probability of success. Using an example in Acute Myeloid Leukemia, this talk aims to illustrate how recent statistical insights and technological resources can be harnessed to map out the design space and highlight locally and globally optimal solutions. Aligning clinical development team priorities to confidently select the best design can ensure speedy and successful execution of the trial.

72. On-demand safety and efficacy insights

[IS 8, (page 6)]

Cathleen JEWELL, *Clinical Analytics,AbbVie*

An interactive analytics platform will be showcased that provides cross-functional scientists and medical monitors direct access to clinical data in intuitive displays and visualizations that enable on-demand insights and agile decision making. These solutions are deployed for early development Oncology clinical trials to promote rapid and repeatable consumption of the data and enable go/no-go decisions. The platform, underlying algorithms, and business process will be highlighted.

73 . mbImpute: an accurate and robust imputation method for microbiome data

[Student Paper Competition 2, (page 20)]

Ruochen JIANG, *Statistics,University of California, Los Angeles*
Vivian Wei LI, *Rutgers School of Public Health*
Jessica Jingyi LI, *University of California, Los Angeles*

Microbiome studies have gained increased attention since many discoveries revealed connections between human microbiome compositions and diseases. A critical challenge in microbiome data analysis is the existence of many non-biological zeros, which distort taxon abundance distributions, complicate data analysis, and jeopardize the reliability of scientific discoveries. To address this issue, we propose the first imputation method for microbiome datambImputeto identify and recover likely non-biological zeros by borrowing information jointly from similar samples, similar taxa, and optional metadata including sample covariates and taxon phylogeny. Comprehensive simulations verify that mbImpute achieves better imputation accuracy under multiple metrics, compared with five state-of-the-art imputation methods designed for non-microbiome data. In real data applications, we demonstrate that mbImpute improves the power of identifying disease-related taxa from microbiome data of type2 diabetes and colorectal cancer, and mbImpute preserves non-zero distributions of taxa abundances.

74 . Improving Exoplanet Detection Power: Multivariate Gaussian Process Models for Stellar Activity

[IS 22, (page 11)]

David JONES, *Department of Statistics,Texas A&M University*
David C. STENNING, *Simon Fraser University*
Eric B. FORD, *Penn State University*
Robert L WOLPERT, *Duke University*
Thomas J LOREDO, *Cornell University*

The radial velocity technique is one of the two main approaches for detecting planets outside our solar system, often referred to as exoplanets. When a planet orbits a star its gravitational force causes the star to move and this induces a Doppler shift (i.e. the star light appears redder or bluer than expected), and it is this effect that the radial velocity method attempts to detect. Unfortunately, these Doppler signals are typically contaminated by various stellar activity phenomena, such as dark spots on the star surface. We propose a Gaussian process modeling framework to capture this stellar activity and thereby improve detection power for low-mass planets (e.g., Earth-like planets). Our approach builds on previous work in two ways: (i) we use dimension reduction techniques to construct data-driven stellar activity proxies, as opposed to using traditional activity proxies; (ii) we extend the multivariate Gaussian process model of Rajpaul et al. (2015) to a class of models and use a large-scale model selection procedure to find the best model for the particular proxies at hand. Our method results in substantially improved power for planet detection compared with existing methods in the astronomy literature.

75. Optimal Product Design by Sequential Experiments in High Dimensions

[IS 50, (page 23)]

Mingyu (Max) JOO, *Marketing,UC Riverside*
Thompson MICHAEL,
Allenby GREG M., *Ohio State University*

The identification of optimal product and package designs is challenged when attributes and their levels interact. Firms recognize this by testing trial products and designs prior to launch where the effects of interactions are revealed. A difficulty in conducting analysis for product design is dealing with the high dimensionality of the design space and the selection of promising product configurations for testing. We propose an experimental criterion for efficiently testing product profiles with high demand potential in sequential experiments. The criterion is based on the expected improvement in market share of a design beyond the current best alternative. We also incorporate a stochastic search variable selection method to selectively estimate relevant interactions among the attributes. A validation experiment confirms that our proposed method leads to improved design concepts in a high-dimensional space compared to alternative methods.

76. Data Splitting

[Special Invited Session 3, (page 25)]

Roshan JOSEPH, *School of Industrial and Systems Engineering,Georgia Institute of Technology*
Akhil VAKAYIL, *Georgia Institute of Technology*

For developing statistical and machine learning models, it is common to split the dataset into two parts: training and testing. The training part is used for fitting the model and the testing part for evaluating the performance of the fitted model. The most common strategy for splitting is to randomly sample a fraction of the dataset. In this talk, I will discuss an optimal method for doing this.

77. Sequential basket trial design based on multi-source exchangeability with predictive probability monitoring

[IS 21, (page 11)]

Alexander KAIZER, *Biostatistics and Informatics,University of Colorado-Anschutz Medical Campus*
Emily ZABOR, *Cleveland Clinic*
Nan CHEN, *Gilead Sciences*
Brian HOBBS, *University of Texas-Austin*

Precision medicine endeavors to conform therapeutic interventions to the individuals being treated and needs to account for the heterogeneity of treatment benefit among patients and patient subpopulations. In oncology, basket trials have emerged as an increasingly popular trial design to better address the goals of precision medicine that endeavors to test the effectiveness of a therapeutic strategy among patients defined by the presence of a particular biomarker target rather than a particular cancer type, where the evaluation of treatment effectiveness are conducted with respect to the ”baskets” which collectively represent a partition of the targeted patient population. However, many basket trials may be incorporating inefficient statistical methodology with respect to approaches for interim monitoring and sharing information across baskets where they may be potentially exchangeable. In this presentation, we will present novel methodology for a sequential basket trial design with Bayesian interim analyses with predictive probability monitoring and the incorporation of a novel hierarchical modeling strategy for sharing information among a collection of discrete, potentially non-exchangeable subtypes that we contrast with the popular, but inefficient, Simon two-stage design.

78. Bayesian Spatial Blind Source Separation via the Thresholded Gaussian Process

[IS 25, (page 12)]

Jian KANG, *Jian Kang, University of Michigan*
Ben WU, *Renmin University*
Ying GUO, *Emory University*

Blind source separation (BSS) aims to separate latent source signals from their mixtures. For spatially dependent signals in high dimensional and large-scale data, such as neuroimaging, most existing BSS methods do not take into account the spatial dependence and the sparsity of the latent source signals. To address these major limitations, we propose a Bayesian spatial blind source separation (BSP-BSS) approach for neuroimaging data analysis. We assume the expectation of the observed images as a linear mixture of multiple sparse and piece-wise smooth latent source signals, for which we construct a new class of Bayesian nonparametric prior models by thresholding Gaussian processes. We assign the von Mises-Fisher priors to mixing coefficients in the model. Under some regularity conditions, we show that the proposed method has several desirable theoretical properties including the large support for the priors, the consistency of joint posterior distribution of the latent source intensity functions and the mixing coefficients and the selection consistency on the number of latent sources. We use extensive simulation studies and an analysis of the resting-state fMRI data in the Autism Brain Imaging Data Exchange (ABIDE) study to demonstrate that BSP-BSS outperforms the existing alternatives for separating latent brain networks and detecting activated brain activation in the latent sources.

79. Applying Double Machine Learning to Targeted Email Promotions: A Journey Down the Conversion Funnel

[IS 53, (page 24)]

Wreetabrata KAR, *Purdue University, Purdue University-Main Campus (West Lafayette, IN)*

We estimate the causal effects of different targeted email promotions on the opening-clicking-purchasing decisions of the consumers who receive them. To do so, we leverage recent advances in causal machine learning techniques to capture heterogeneity in the content of the email subject line itself, as well as heterogenous consumer responses to the promotional offers and semantic cues contained therein. We make two primary contributions to the literature. First, we demonstrate how machine learning can be leveraged in a two-step framework to provide unbiased estimates of the effect of separate components in a high-dimensional marketing intervention exploiting purely observational data. Second, we apply our methodology to data on 33 distinct email promotions sent by a retailer to nearly 2 million individuals on their contact list to highlight quantitative findings that are relevant to both the marketing literature and managerial practice.

80. Reinforced designs: Multiple instruments plus control groups as evidence factors in an observational study

[IS 4, (page 4)]

Bikram KARMAKAR, *Statistics, University of Florida*
Dylan S. SMALL, *University of Pennsylvania*
Paul R. ROSENBAUM, *University of Pennsylvania*

A series of observational studies can replicate each other in finding an association between an exposure and an outcome. But if each of these studies is susceptible to unmeasured biases in the same way this replication does not strengthen the evidence for a treatment effect. Conducting the same observational study with a larger sample size also does not strengthen the evidence. To be of value, a replication should remove, or reduce, or at least vary a potential source of bias that resulted in uncertainty in earlier studies. Multiple analyses provide evidence about unmeasured biases if: (i) certain biases that would invalidate one analysis do not bias another analysis, (ii) each analysis is insensitive to small or moderate biases of the type that might invalidate that analysis, and (iii) these several analyses would be nearly statistically independent if the treatment had no effect. Analyses of this type are possible in a single study; these analyses are called evidence factors. I will provide a general introduction to evidence factors. Evidence factors analysis is often available to us in various types of studies. I will present some design and analytic tools to form evidence factors in a study. I will demonstrate these tools in forming 3 evidence factors in a study of the effectiveness of Catholic high schools on income that has a couple of instruments: being Catholic and living close to a Catholic school; neither of which is unarguably a valid instrument for going to a Catholic school.

81. Innovative Statistical Thinking in Support of Translational Science and Companion Diagnostic Development

[IS 38, (page 18)]

Maha KARNOUB, *Maha Karnoub, Daiichi Sankyo -*

Translational Medicine Biostatistics

Statistical thinking is integral in the research and decision-making process. The discipline is framed by logic and the system through which hypotheses can be developed or tested. Statistical training provides expertise in implementing the optimal designs and applying the appropriate analyses techniques to achieve this objective. Therefore, beyond being skilled in the techniques, an innovative statistician is able to participate in the synthesis of information discussed with the project teams, and to provide timely feedback on the follow-on experiment to be developed or executed to answer the questions from the team. This requires a thorough understanding of the various potential contributions from the various members of the analytical team. Translational Medicine statisticians are at the center of a collaborative model, taking on guidance from the Project Team, from Translational Sciences, and from Biomarker and Companion Diagnostics leads, and efficiently teaming with Data Management, Programming and Bioinformatics, to carry through the research and diagnostics objectives at Daiichi-Sankyo. We will describe some of the initiatives we setup since developing the Translational Medicine Biostatistics group a little over a year ago, and the types of projects we supported using this model.

82. X-ray Astronomy in 4 Dimensions

[IS 22, (page 11)]

Vinay KASHYAP, *High-Energy Astrophysics Division, Center for Astrophysics – Harvard & Smithsonian*
CHASC CHASC, *CHASC*

We count individual photons in high-energy astronomy, which means each photon can be tagged with several properties the direction it came from, the time it arrived, and the energy it deposited. Our CHASC Astrostatistics group has developed several methods to jointly analyze such high-dimensional data, and I will describe some of them. Specifically, I will discuss measuring boundaries of extended sources in 2D location space, and disambiguation of overlapping sources in location-energy-time space.

83 . Lag selection and estimation in mixed frequency regression using Bayesian nested lasso

[IS 64, (page 29)]

Kshitij KHARE, *Department of Statistics, University of Florida*
Satyajit GHOSH, *FDA*
George MICHAILIDIS, *University of Florida*

Even though many time series are sampled at different frequencies, their joint evolution is usually modeled and analyzed at a common low frequency. The Mixed Data Sampling (MIDAS) framework was developed to enable joint modeling of mixed frequency temporally evolving data. In this paper, we develop a fully Bayesian method to jointly estimate both selection of the appropriate lag, as well as the regression coefficients in linear models wherein the response is measured at lower frequency than the predictors. This is accomplished through a novel prior distribution, coined the Bayesian Nested Lasso (BNL), that leads to principled selection of the lag of the predictors, reduces the effective number of model parameters through sparsity induced by the lasso component and finally incorporates desirable decay patterns over time lags in the magnitude of the corresponding regression coefficients. Further, it is easy to obtain samples from the posterior distribution due to the closed form expressions for the conditional distributions of the model parameters. Numerical results obtained from synthetic data illustrate the good performance of the proposed Bayesian framework, both in parameter selection and estimation, as well as in forecasting.

84 . Efficient competing risks regression models under the generalized case-cohort design

[IS 24, (page 12)]

Soyoung KIM, *Division of Biostatistics,Medical College of Wisconsin*
Yayun XU, *Merck Pharmaceutical company,*
Mei-Jie Zhang, Kwang Woo AHN, *Division of Biostatistics, Medical College of Wisconsin*
David COUPER, *Department of Biostatistics, University of North Carolina at Chapel Hill*

A generalized case-cohort design has been used when measuring exposures is expensive and events are not rare in the full cohort. This design collects expensive exposure information from a (stratified) randomly selected subset from the full cohort, called the subcohort, and a fraction of cases outside the subcohort. For the full cohort study with competing risks, He et al. (2016) studied the non-stratified proportional subdistribution hazards model with covariate-dependent censoring to directly evaluate covariate ef-

fects on the cumulative incidence function. In this paper, we propose a stratified proportional subdistribution hazards model with covariate-adjusted censoring weights for competing risks data under the generalized case-cohort design. We consider a general class of weight functions to account for the generalized case-cohort design. Then, we derive the optimal weight function which minimizes the asymptotic variance of parameter estimates within the general class of weight functions. The proposed estimator is shown to be consistent and asymptotically normally distributed. The simulation studies show (i) the proposed estimator with covariate-adjusted weight is unbiased when the censoring distribution depends on covariates; and (ii) the proposed estimator with the optimal weight function gains parameter estimation efficiency. We apply the proposed method to stem cell transplantation and diabetes data sets.

85. Nonparametric Variable Screening with Decision Stumps

[IS 59, (page 27)]

Jason KLUSOWSKI, *Operations Research and Financial Engineering,Princeton University*
Peter M. TIAN, *Princeton University*

Decision trees and their ensembles are endowed with a rich set of diagnostic tools for ranking and screening variables in a predictive model. Despite the widespread use of tree based variable importance measures, pinning down their theoretical properties has been challenging and therefore largely unexplored. To address this gap between theory and practice, we derive finite sample performance guarantees for variable selection in nonparametric models using a single-level CART decision tree (a decision stump). Under standard operating assumptions in variable screening literature, we find that the marginal signal strength of each variable and ambient dimensionality can be considerably weaker and higher, respectively, than state-of-the-art nonparametric variable selection methods. Furthermore, unlike previous approaches that attempt to directly estimate each marginal projection via a truncated basis expansion, the fitted model used here is a simple, parsimonious decision stump, thereby eliminating the need for tuning the number of basis terms.

86. Inference for Differential Networks in a High-dimensional Setting

[IS 28, (page 13)]

Mladen KOLAR, *The University of Chicago Booth School of Business,The University of Chicago Booth School of Business*
Byol KIM, *Irina Gaynanova*

We consider the problem of constructing confidence intervals for the differential edges between the two high-dimensional networks. The problem is motivated by the comparison of gene interactions between two molecular subtypes of colorectal cancer with distinct survival prognosis. Unlike the existing approaches for differential network inference that require sparsity of individual precision matrices from both groups, we only require sparsity of the precision matrix difference. We discuss the methods' theoretical properties, evaluate its performance in numerical studies and highlight directions for future research.

87. PROFIT: Projection-based Test in Longitudinal Functional Data

[Student Paper Competition 1, (page 15)]

Koner SALIL, *North Carolina State University*
So Young PARK,
Ana-Maria STAICU,

In many modern applications, a dependent functional response is observed for each subject over repeated time, leading to longitudinal functional data. In this paper, we propose a novel statistical procedure to test whether the mean function varies over time. Our approach relies on reducing the dimension of the response using data-driven orthogonal projections and it employs a likelihood-based hypothesis testing. We investigate the methodology theoretically and discuss a computationally efficient implementation. The proposed test maintains the type I error rate, and exhibits adequate power to detect the departure from the null hypothesis in finite sample simulation studies. We apply our method to the longitudinal diffusion tensor imaging study of multiple sclerosis (MS) patients to formally assess whether the brain's health tissue, as summarized by fractional anisotropy (FA) profile, changes over time during the study period.

88. High-dimensional rank-based inference for testing relative effects

[IS 39, (page 18)]

Xiaoli KONG, *Department of Mathematics and Statistics,Loyola University Chicago*

Solomon HARRAR, *University of Kentucky*

In this talk, a fully nonparametric rank-based method is introduced for comparing multi-group relative effects. No assumption has been made on the distribution. We only require that the dependencies between the variables satisfy some mild conditions. In particular, to develop the theory we prove a novel result for studying the asymptotic behavior of quadratic forms in ranks. The simulation results show 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 results are applied to Electroencephalograph (EEG) data that arose from a study to examine the correlates of genetic predisposition to alcoholism.

89. Refined Maximal Inequalities with Applications to Oracle Inequalities

[IS 5, (page 4)]

Arun KUCHIBHOTLA, *Department of Statistics and Data Science,Carnegie Mellon University*

In this talk, I will discuss refined maximal inequalities for a maximum of averages of heavy-tailed random vectors. These will be used to derive maximal inequalities for empirical process. Applications to derive rates of convergence, as well as oracle inequalities, for the LSE with heavy-tailed errors will also be discussed.

90. Composite responder rate estimation under non-ignorable missingness

[IS 49, (page 22)]

Madan KUNDU, *Daiichi Sankyo Inc*

Missing data are a potential source of bias in clinical trials including the trials evaluating (composite) responder rate based on one or more measurements at a particular visit. Composite response criteria are useful when therapeutic benefit is determined by simultaneous improvement on more than one characteristic and are in use in many therapeutic areas including irritable bowel syndrome, myelofibrosis and rheumatoid arthritis. The drug regulatory agencies often recommend adopting a missing-not-at-random (MNAR) mechanism for handling missing data. Existing likelihood based MNAR missing data methods (e.g., selection model or pattern mixture model) make some distributional assumption such as multivariate normal across visits and the resulting conclusion can be very sensitive to this assumption. Therefore, we propose an intuitive and straightforward strategy to estimate the responder rate under MNAR missingness without making any distributional assumption. By not using distributional assumptions our method avoids bias due to mis-specification of distribution across visits. Our method falls in the class of inverse-probability-weighted (IPW) estimation. However, unlike IPW approach, the proposed approach proceeds by estimating the responder rates separately among completers and dropped-out patients rather than computing weights for individual completers. Our method arguably is more intuitive and requires less technical sophistication compared to existing approaches and is also less computationally intensive. The comparative performance of our proposed estimator was evaluated through simulation and the method was applied to clinical trial data comparing composite responder rates of two treatments.

91. Semi-parametric Bayes Regression and Variable Selection Using Network Valued Covariates.

[IS 42, (page 19)]

Suprateek KUNDU, *Biostatistics,Emory University*
Xin MA, *Emory University*
Jennifer STEVENS, *Emory University*

There is an increasing recognition of the role of brain networks as neuroimaging biomarkers in mental health and psychiatric studies. Our focus is post-traumatic stress disorder (PTSD), where the brain network interacts with environmental exposures in complex ways to drive the disease progression. Existing linear models seeking to characterize the relation between the clinical phenotype and the entire edge set in the brain network may be overly simplistic and often involve inflated number of parameters leading to computational burden and inaccurate estimation. In one of the first such efforts, we develop a novel two stage Bayesian framework to find a node-specific lower dimensional representation for the network using a latent scale approach in the first stage, and then use a flexible Gaussian process regression framework for prediction involving the latent scales and other supplementary covariates in the second stage. The proposed approach relaxes linearity assumptions, addresses the curse of dimensionality and is scalable to high dimensional networks while maintaining interpretability at the node level of the network. Extensive simulations and results from our motivating PTSD application show a distinct advantage of the proposed approach over competing linear and nonlinear approaches in terms of prediction and coverage.

92 . Modeling Heterogeneity in Consumer Preferences using Bayesian Methods

[IS 41, (page 19)]

Choudur LAKSHMINARAYAN, *Teradata Labs,Teradata Labs*

In Marketing applications, the availability of a galore of brands and products provides a dazzling array of choices to consumers. Therefore, modeling consumer preferences in the aggregate across brands and products is insufficient and inefficient. The heterogeneity in consumer choices renders marketing communications and advertising very challenging. Marketing departments are well served by statistical models that produce product level parameter estimates and statistical inferences. It is therefore incumbent to not only obtain the point estimates of the product level parameters, but also the associated uncertainty in these estimates by obtaining posterior distributions of the parameters under heterogeneity invoking Bayesian hierarchical models.

93. A Joint Spatial Conditional Auto-Regressive Model for Estimating HIV Prevalence Rates Among Key Populations

[IS 66, (page 29)]

Zhou LAN, *Yale School of Medicine,Yale School of Medicine*
Le BAO, *Penn State University*

Ending the HIV/AIDS pandemic is among the Sustainable Development Goals for the next decade. In order to overcome the gap between the need for care and the available resources, better understanding of HIV epidemics is needed to guide policy decisions, especially for key populations that are at higher risk for HIV infection. Accurate HIV epidemic estimates for key populations have been difficult to obtain because their HIV surveillance data is very limited. In this paper, we propose a so-called joint spatial conditional auto-regressive model for estimating HIV prevalence rates among key populations. Our model borrows information from both neighboring locations and dependent populations. As illustrated in the real data analysis, it provides more accurate estimates than independently fitting the sub-epidemic for each key population. In addition, we provide a study to reveal the conditions that our proposal gives a better prediction. The study combines both theoretical investigation and numerical study, revealing strength and limitations of our proposal.

94 . Some recent models using binary tree ensembles for various outcome types

[IS 58, (page 26)]

Purushottam LAUD, *Medical College of Wisconsin*
Robert MCCULLOCH, *Arizona State University*
Rodney SPARAPANI, *Medical College of Wisconsin*
Brent LOGAN, *Medical College of Wisconsin*

Bayesian additive regression trees (BART) is a binary tree ensemble method that allows for high-dimensional flexible regression surfaces in the model, built via set-wise constant functions on grids generated by ensembles of binary trees. The Bayesian approach is particularly well suited, via suitable prior specifications, to generate large flexible collections of regression surfaces. At the same time, using standard Bayesian updating and modern Markov chain sampling techniques, inference and prediction conditional on data is readily implementable. Since the seminal paper by Chipman, George and McCulloch (2010 Annals of Applied Statistics), this nonparametric regression technique has seen much development and extension to address various data types and modeling purposes. We will present a selection of these - mainly recent and emerging models - that showcase the versatility of the basic BART structure.

95 . Comparative Biomarker Modeling for Optimal Patient Selection in Immuno-oncology

[IS 38, (page 18)]

Jae LEE, *Biomarker Group, Global Biometric and Data Sciences,Bristol Myers Squibb*

The study presents an example of composite biomarker modeling for optimal treatment allocation between IO treatments. A fuller abstract will be updated after an internal review.

96. On weighted log-rank combination tests and companion Cox model estimators

[IS 61, (page 28)]

Larry LEON, *Biostatistics,Bristol-Myers Squibb*
Ray LIN, *Genentech*

Keaven ANDERSON, *Merck*

In randomized clinical trials the log-rank test and Cox proportional hazards model are the gold standard in survival data analyses. While the log-rank test is generally valid, in the presence of non-proportional hazards the power can be substantially decreased relative to the proportional hazards assumptions under which studies are usually designed. In contrast, weighted log-rank tests can be more powerful for specific treatment differences under non-proportional hazards scenarios; However, a poor choice of the weighting form can be detrimental. Recent work on combining various weighted log-rank tests allows for tests that are capable of detecting treatment effects across a broad range of non-proportional hazards scenarios. In this talk we discuss these ideas with a framework based on a flexible resampling approach which allows for the combination of various testing procedures in addition to weighted log-rank tests. In particular, we describe how tests based on restricted mean survival time (RMST) comparisons can be included within combinations of weighted log-rank tests. For estimation, we propose companion weighted Cox model estimators (Lin, 1991; Sasieni, 1993a,b) which utilize the weighting form that is selected through the combination test and provide simultaneous confidence intervals. The performance of various combinations and their companion Cox estimators as well as RMST are evaluated in simulation studies under null, proportional hazards, late-separation, and early-separation scenarios.

97. Nonparametric Bayesian Analysis of Genotoxicity Screening Assays

[IS 52, (page 23)]

Dingzhou (Dean) LI, *Drug Safety Statistics, Pfizer,Drug Safety Statistics, Pfizer*

Early classification of genotoxicity plays an important role in compound screening and drug development. High throughput, high content in vitro screening assays are used for this purpose; however, high false positive rates may arise from using traditional statistical methods. Also, the variety of assays makes it hard to transform the data to normality. This presentation reviews some nonparametric Bayesian methods for positivity identification using multi-plate data that contain both historical negative controls and positive controls.

98. Linear regression and its inference on noisy network-linked data

[IS 54, (page 24)]

Tianxi LI, *Department of Statistics,University of Virginia*
Can M. LE, *UC Davis*

Linear regression on a set of observations linked by a network has been an essential tool in modeling the relationship between response and covariates with additional network data. Despite its wide range of applications in many areas, such as in social sciences and health-related research, the problem has not been well-studied in statistics so far. Previous methods either lack inference tools or rely on restrictive assumptions on social effects and usually assume that networks are observed without errors, which is unrealistic in many problems. This talk proposes a linear regression model with nonparametric network effects. The model does not assume that the relational data or network structure is exactly observed; thus, the method can be provably robust to a certain network perturbation level. A set of asymptotic inference results is established under a general requirement of the network observational errors, and the robustness of this method is studied in the specific setting when the errors come from random network models. We discover a phase-transition phenomenon of the inference validity concerning the network density when no prior knowledge of the network model is available while also showing a significant improvement achieved by knowing the network model. As a by-product of this analysis, we derive a rate-optimal concentration bound for random subspace projection that may be of independent interest. Extensive simulation studies are conducted to verify these theoretical results and demonstrate the advantage of the proposed method over existing work in terms of accuracy and computational efficiency under different data-generating models. The method is then applied to adolescent network data to study the gender and racial difference in social activities.

99. Mixture cure rate models with BART

[IS 15, (page 9)]

Xiao LI, *Biostatistics,Medical College of Wisconsin*
Rodney SPARAPANI, *Medical College of Wisconsin*
Brent LOGAN, *Medical College of Wisconsin*

In many clinical studies with survival outcomes, a fraction of patients may be considered cured by the therapy in terms of achieving long-term survival or disease control. Such data can be analyzed using mixture cure rate models that assume the population contains a mixture of cured and uncured patients. Cure status is a latent variable for those censored patients. Recently, flexible survival prediction models using machine learning techniques have been used to improve predictive performance while minimizing assumptions about the functional form of the relationship between covariates and outcomes. We propose an extension to a Bayesian machine learning technique called Bayesian Additive Regression Trees (BART) to address mixture cure models. Two BART models are used to simultaneously model the cure probability, a probit BART; and the survival distribution in uncured individuals via a log-normal accelerated failure time BART model. The performance of the BART mixture cure rate model is evaluated on simulated data. We also apply our model to a real study of patients undergoing a potentially curative therapy of hematopoietic stem cell transplantation and compare the results with others.

100. Consistency of Spectral Clustering on Hierarchical Stochastic Block Models

[IS 9, (page 6)]

Xiaodong LI, *UC Davis/Statistics,UC Davis*
Lihua LEI, *Stanford University*
Xingmei LOU, *UC Davis*

We apply a generic network model, based on the Stochastic Block Model, to study the hierarchy of communities in real-world networks, under which the connection probabilities are structured in a binary tree. Under the network model, we show that the eigenstructure of the expected unnormalized graph Laplacian reveals the community structure of the network as well as the hierarchy of communities in a recursive fashion. Inspired by the nice property of the population eigenstructure, we develop a recursive bipartitioning algorithm that divides the network into two communities based on the Fiedler vector of the unnormalized graph Laplacian and repeats the split until a stopping rule indicates no further community structures. We prove the weak and strong consistency of our algorithm for sparse networks with the expected node degree in $O(logn)$ order, based on newly developed theory on $L_{2,infty}$ eigenspace perturbation, without knowing the total number of communities in advance. Unlike most of existing work, our theory covers multi-scale networks where the connection probabilities may differ in order of magnitude, which comprise an important class of models that are practically relevant but technically challenging to deal with. Finally we demonstrate the performance of our algorithm on synthetic data and real-world examples.

101. Rerandomization and Regression Adjustment

[IS 56, (page 26)]

Xinran LI, *Department of Statistics,University of Illinois at Urbana-Champaign*
Peng DING, *University of California, Berkeley*

Randomization is a basis for the statistical inference of treatment effects without strong assumptions on the outcome-generating process. Appropriately using covariates further yields more precise estimators in randomized experiments. R. A. Fisher suggested blocking on discrete covariates in the design stage or conducting analysis of covariance (ANCOVA) in the analysis stage. We can embed blocking into a wider class of experimental design called rerandomization, and extend the classical ANCOVA to more general regression adjustment. Rerandomization trumps complete randomization in the design stage, and regression adjustment trumps the simple difference-in-means estimator in the analysis stage. It is then intuitive to use both rerandomization and regression adjustment. Under the randomization-inference framework, we establish a unified theory allowing the designer and analyzer to have access to different sets of covariates. We find that asymptotically (a) for any given estimator with or without regression adjustment, rerandomization never hurts either the sampling precision or the estimated precision, and (b) for any given design with or without rerandomization, our regression-adjusted estimator never hurts the estimated precision. Therefore, combining rerandomization and regression adjustment yields better coverage properties and thus improves statistical inference. To theoretically quantify these statements, we discuss optimal regression-adjusted estimators in terms of the sampling precision and the estimated precision, and then measure the additional gains of the designer and the analyzer. We finally suggest using rerandomization in the design and regression adjustment in the analysis followed by the Huber–White robust standard error.

102. Bayesian Regularization and Estimation for Gaussian Conditional Ran-

dom Fields

[IS 63, (page 28)]

Feng LIANG, *Statistics Department, UIUC,University of Illinois at Urbana-Champaign*
Lingrui GAN, *Facebook*
Naveen N. NARISETTY, *University of Illinois at Urbana-Champaign*

To address overfitting, a central issue in statistics and machine learning, many successful techniques have been formulated under this mathematical framework known as regularization, which penalizes or adds constraints on the underlying model parameters. In this talk, we introduce a general framework for effective regularization from a Bayesian perspective.

We illustrate the application of our general framework to estimating a Gaussian conditional random field model, through which we can learn the conditional dependence structures among multiple outcomes, and between the outcomes and a set of covariates simultaneously. We investigate the corresponding MAP (maximum a posteriori) estimators that require dealing with a non-convex optimization problem. In spite of the non-convexity, we establish statistical accuracy for all points in the high posterior density region and some local optima. For fast and efficient computation, an EM algorithm is proposed to compute the MAP estimator of the parameter and (approximate) posterior probabilities on the edges of the underlying sparse structure. Through simulation studies and a real application, we have demonstrated the fine performance of our method compared with existing alternatives.

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

[IS 46, (page 21)]

Ruitao LIN, *Ruitao Lin,The University of Texas MD Anderson Cancer Center*

For immunotherapy, such as checkpoint inhibitors and chimeric antigen receptor T-cell therapy, where the efficacy does not necessarily increase with the dose, the maximum tolerated dose may not be the optimal dose for treating patients. For these novel therapies, the objective of dose-finding trials is to identify the optimal biologic dose (OBD) that optimizes patients risk-benefit trade-off. We propose a simple and flexible Bayesian optimal interval phase I/II (BOIN12) trial design to find the OBD that optimizes the risk-benefit trade-off. 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 trade-off. We performed simulation studies to evaluate the performance of the BOIN12 design. Compared with 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 pretabulated 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. The BOIN12 design is simple to implement and yields desirable operating characteristics. It overcomes the computational and implementation complexity that plagues existing Bayesian phase I/II dose-finding designs and provides a useful design to optimize the dose of immunotherapy and targeted therapy. User-friendly software is freely available to facilitate the application of the BOIN12 design.

104. SBL: Bayesian Lasso for detecting haplotypes associated with survival traits

[IS 37, (page 17)]

Shili LIN, *Statistics,Ohio State University*

Rare genetic variants are a key factor in understanding the etiology of common diseases. Although the literature has primarily focused on rare single nucleotide variants (rSNVs), rare haplotype variants (rHTVs) offer perhaps even greater biological relevance, as variants on the same chromosomes are often passed jointly. Several methods have been developed to detect rHTV effects on common diseases based on the Bayesian Lasso methodology for binary and quantitative traits (LBL), and greater power over a number of rSNV based methods has been demonstrated. In this talk, I will describe a Bayesian Lasso method for detecting rHTVs associated with survival traits, referred to as Survival Bayesian Lasso (SBL). Simulation studies and an analysis of The Cancer Genome Atlas breast cancer data will be presented to demonstrate the utility of SBL.

105. Scalable Integrative Analysis of Large Genome and Phenome Data

[Plenary Lecture 3, (page 25)]

Xihong LIN, *Departments of Biostatistics and Department of Statistics, Harvard University and Broad Institute*

Big data from genome, exposome, and phenome are becoming available at a rapidly increasing rate with no apparent end in sight. Examples include Whole Genome Sequencing data, smartphone data, wearable devices, and Electronic Health Records (EHRs). A rapidly increasing number of large scale national and institutional biobanks have emerged worldwide. Biobanks integrate genotype, electronic health records, and lifestyle data, and is the trend of health science research. In this talk, I discuss opportunities, analytic tools and resources, and challenges presented by large scale biobanks and population-based Whole Genome Sequencing (WGS) studies of common and rare genetic variants and EHRs of many phenotypes. I will also discuss integrative analysis of different types of data within the causal mediation analysis framework. The discussions are illustrated using ongoing large scale whole genome sequencing studies of the Genome Sequencing Program of the National Human Genome Research Institute and the Trans-Omics Precision Medicine Program from the National Heart, Lung and Blood Institute, and the UK Biobank and FinnGen.

106. Causal inference and estimands in clinical trials

[IS 1, (page 3)]

Ilya LIPKOVICH, *Real world analytics,Eli Lilly and Company*
Bohdana RATITCH, *Bayer*
Craig MALLINCKRODT, *Biogen Idec*

The National Research Council (2010) report on the prevention and treatment of missing data highlighted the need to clearly specify causal estimands. The ICH E9(R1) addendum (2019) was another major step in promoting the use of the causal estimands framework in clinical practice. The language of potential outcomes (PO) is widely accepted in the causal inference literature but is not yet recognized in the clinical trial community and was not used in defining causal estimands in ICH E9(R1). We argue that the use of PO language and solid causal foundation in our thinking and writing can help to further disambiguate estimand definitions. In this presentation, we bridge the gap between the causal inference community and clinical trialists by advancing the use of causal estimands in clinical trial settings. We illustrate how concepts from causal literature, such as POs and dynamic treatment regimens, can facilitate defining and implementing causal estimands for different types of outcomes providing a unifying language for both observational and randomized clinical trials.

107. Design and Challenges in Platform Design in the Immuno-Oncology Drug Development with application in NSCLC

[IS 55, (page 25)]

Feng LIU, *Oncology Biometrics,AstraZeneca*

A platform master protocol design refers to those study design in which multiple treatments or treatment regimens are tested simultaneously and the platform design offers flexible features such as discontinuing treatments earlier for futility or superiority, graduation to next phase of study, or adding new treatment regime to be tested under one overarching master protocol. The design features of the platform design with application in a Phase 2 NSCLC using immune-oncology agents will be discussed with primary endpoint as OS. Design challenges such as statistical considerations in handling contemporary and non-contemporary control as well as regulatory interactions will also be discussed as well as developmental challenges highlighted including analysis consideration of primary endpoint under non-proportional hazard, development of stopping and graduation criteria, treatment agnostic protocol element and drug specific amendments, and challenge in randomization and drug supply, etc. Due to the design complexity, cross functions or cross industry collaborative efforts are critical in the development and execution of the platform design.

108 . Fine-tuning genetic prediction models using marginal association statistics

[IS 12, (page 7)]

Qiongshi LU, *Department of Biostatistics and Medical Informatics,University of Wisconsin-Madison*

Polygenic risk scores (PRSs) have wide applications in human genetics research. Notably, most PRS models include tuning parameters which improve predictive performance when properly selected. However, existing model-tuning methods require individual-level genetic data as the training dataset or as a validation dataset independent from both training and testing samples. These data rarely exist in practice, creating a significant gap between PRS methodology and applications. Here, we introduce PUMAS (Parameter-tuning Using Marginal Association Statistics), a novel method to fine-tune PRS models using summary statistics from genome-wide association studies (GWASs). Through extensive simulations, external validations, and analysis of 65 traits, we demonstrate that PUMAS can perform a variety of model-tuning procedures (e.g. cross-validation) using GWAS summary statistics and can effectively benchmark and optimize PRS models under diverse genetic architecture. On average, PUMAS improves the predictive R2 by 205.6% and 62.5% compared to PRSs with arbitrary p-value cutoffs of 0.01 and 1, respectively. Applied to 211 neuroimaging traits and Alzheimers disease, we show that fine-tuned PRSs will significantly improve statistical power in downstream association analysis. We believe our method resolves a fundamental problem without a current solution and will greatly benefit genetic prediction applications.

109 . Statistical Computing Meets Quantum Computing

[IS 3, (page 4)]

Ping MA, *Statistics, University of Georgia*

With the rapid development of quantum computers, quantum computing has been studied extensively. Unlike electronic computers, a quantum computer operates on quantum processing units, or qubits, which can take values 0, 1, or both simultaneously due to the superposition property. The number of complex numbers required to characterize quantum states usually grows exponentially with the size of the system. For example, a quantum system with p qubits can be in any superposition of 2^p orthonormal states simultaneously, while a classical system can only be in one state at a time. Such a paradigm change has motivated significant developments of scalable quantum algorithms in many areas.

However, quantum algorithms tackling statistical problems are still lacking. In this talk, I will present challenges and opportunities for developing quantum algorithms. I will introduce a novel quantum algorithm for the best subset selection problem.

110. A Negative Binomial Mixed Effects Location-Scale Model for Physical Activity Data Provided by Wearable Devices

[IS 60, (page 27)]

Qianheng MA, *Department of Public Health Sciences, University of Chicago*

Genevieve F. DUNTON, *University of Southern California*

Donald HEDEKER, *University of Chicago*

In recent years, the use of wearable devices, e.g., actigraphies, have become increasingly prevalent. Wearable devices enable more accurate real-time tracking of a subject's physical activity (PA) level, such as step counts or minutes in moderate-to-vigorous intensity PA (MVPA), which are important general health markers. The intensive within-subject data provided by wearable devices, e.g., minutes in MVPA summarized per hour across days and even months, allow the possibility for modeling not only the mean PA level, but also the dispersion level for each subject. Especially in the context of daily physical activity, subjects' dispersion levels are potentially informative in reflecting their exercise patterns: some subjects might exhibit consistent PA across time and can be considered "less dispersed" subjects; while others might have a large amount of PA at a particular time point, while being sedentary (PA=0) for most of the day, and can be considered "more dispersed" subjects. Thus, we propose a negative binomial mixed effects location-scale model to model these intensive longitudinal PA counts and to account for the heterogeneity in both the mean and dispersion level across subjects. Further, to handle the issue of inflated numbers of zeros in the PA data, we also proposed a hurdle/zero-inflated version which additionally includes the modeling of the probability of having ¿ 0 PA levels.

111. Quantile trend filtering

[IS 47, (page 22)]

OSCAR HERNAN MADRID PADILLA, *Statistics, University of California, Los Angeles*

In this talk, we will discuss quantile trend filtering, a recently proposed method for nonparametric quantile regression with the goal of generalizing existing risk bounds known for the usual trend filtering estimators which perform mean regression. We study both the penalized and the constrained version (of order $r \geq 1$) of quantile trend filtering. Our results show that both the constrained and the penalized version (of order $r \geq 1$) attain the minimax rate up to log factors, when the $(r-1)$th discrete derivative of the true vector of quantiles belongs to the class of bounded variation signals. We also show that if the true vector of quantiles is a discrete spline with a few polynomial pieces then the constrained version attains a near parametric rate of convergence. Corresponding results for the usual trend filtering estimators are known to hold only when the errors are sub-Gaussian. In contrast, our risk bounds are shown to hold under minimal assumptions on the error variables. In particular, no moment assumptions are needed and our results hold under heavy-tailed errors. On the other hand, we prove all our results for a Huber type loss which can be smaller than the mean squared error loss employed for showing risk bounds for usual trend filtering.

112. Building an external control arm for development of a new molecular entity

[IS 30, (page 14)]

Antara MAJUMDAR, *Statistical Innovations Group,Medidata Acorn AI*

Ruthie DAVI, *Medidata Acorn AI*

In certain indications, it is well understood that randomized controlled trials lead to slow enrollment and high differential drop-out rate in the standard of care control arm when the standard of care is undesirable to patients. This not only impacts the pace of drug development but may also render a randomized trial uninterpretable if drop-out from the control arm is common. Single arm trials are common in such indications. An external control arm (ECA) built from subject-level data using a propensity score method (Rosenbaum and Rubin, 1983) from subjects outside the current trial but who meet the same eligibility criteria as the subjects of the current trial, is valuable in assessing treatment effect of a new drug that cannot be otherwise assessed with a single arm trial, or an unintentionally under-powered randomized controlled trial. We will present our experience with building an ECA for drug development related to regulatory activities associated with a new molecular entity.

113. Group Orthogonal Supersaturated Designs

[IS 50, (page 23)]

Dibyen MAJUMDAR, *Math., Stat. and Comp., Sci, UIC,University of Illinois at Chicago*

We consider screening experiments where an investigator wishes to study many factors using fewer observations. Research on design of these experiments effectively started with the paper of Booth and Cox (1962) where the authors introduced the concept of E(s2) designs. Since then, there has been considerable research in the area, with the principle focus on main effects models with factors at 2 levels each. In this talk we will briefly discuss E(s2) and UE(s2) optimality and introduce the method for constructing supersaturated designs that is based on the Kronecker product of two matrices. This construction method leads to a partitioning of the factors of the design such that the factors within a group are correlated to the others within the same group but are orthogonal to any factor in any other group. The resulting designs are group-orthogonal supersaturated designs.

114. Simultaneous Selection of Multiple Important Single Nucleotide Polymorphisms in Familial Genome Wide Association Studies Data

[IS 41, (page 19)]

Subhabrata MAJUMDAR, *Data science and AI Research,University of Minnesota*

Saonli BASU, *University of Minnesota*

Matt MCGUE, *University of Minnesota*

Snigdhansu CHATTERJEE, *University of Minnesota*

We propose a resampling-based fast variable selection technique for selecting important Single Nucleotide Polymorphisms (SNP) in multi-marker mixed effect models used in twin studies. Due to computational complexity, current practice includes testing the effect of one SNP at a time, commonly termed as ‘single SNP association analysis’. Joint modeling of genetic variants within a gene or pathway may have better power to detect the relevant genetic variants, hence we adapt our recently proposed framework of e-values to address this. In this paper, we propose a computationally efficient approach for single SNP detection in families while utilizing information on multiple SNPs simultaneously. We achieve this through improvements in two aspects. First, unlike other model selection techniques, our method only requires training a model with all possible predictors. Second, we utilize a fast and scalable bootstrap procedure that only requires Monte-Carlo sampling to obtain bootstrapped copies of the estimated vector of coefficients. Using this bootstrap sample, we obtain the e-value for each SNP, and select SNPs having e-values below a threshold. We illustrate through numerical studies that our method is more effective in detecting SNPs associated with a trait than either single-marker analysis using family data or model selection methods that ignore the familial dependency structure. We also use the e-values to perform gene-level analysis in nuclear families and detect several SNPs that have been implicated to be associated with alcohol consumption.

115. Supervised compression of data

[IS 57, (page 26)]

Simon MAK, *Statistical Science,Duke University*
V. Roshan JOSEPH, *Georgia Institute of Technology*

The phenomenon of big data has become ubiquitous in nearly all disciplines, from science to engineering. A key challenge is the use of such data for fitting statistical and machine learning models, which can incur high computational and storage costs. One solution is to perform model fitting on a carefully-selected subset of the data. Various data reduction methods have been proposed in the literature, ranging from random subsampling to optimal experimental design-based methods. However, when the goal is to learn the underlying input-output relationship, such reduction methods may not be ideal, since it does not make use of information contained in the output. To this end, we propose a supervised data compression method called supercompress, which integrates output information by sampling data from regions most important for modeling the desired input-output relationship. An advantage of supercompress is that it is nonparametric – the compression method does not rely on parametric modeling assumptions between inputs and output. As a result, the proposed method is robust to a wide range of modeling choices. We demonstrate the usefulness of supercompress over existing data reduction methods, in both simulations and a taxicab predictive modeling application.

116. Bayesian local models using partitions

[Special Invited Session 3, (page 25)]

Bani MALLICK, *Statistics,Texas A&M*

All models are not wrong, at least some of them could be correct locally! Moreover, they are useful! Based on this principle I will propose Bayesian local models using partitioning. The Bayesian partition model constructs arbitrarily complex models by splitting the covariate space into an unknown number of disjoint regions. Within each region the data are assumed to be generated by a simpler model. The partition can be created using Voronoi Tessellations or Trees. The main challenge is to determine the local regions (partitions) adaptively. I will mainly discuss the local conditional density regression in this framework. I will also describe the extension of this method for survival and spatial models. Some theoretical properties of the models will be discussed. I will show simulations and applications to real data results where the method will successfully estimate the partition structure as well as the local model parameters.

117. Differential expression of single-cell RNA-seq data using Tweedie models

[IS 18, (page 10)]

Himel MALLICK, *Biostatistics and Research Decision Sciences,Merck Research Laboratories*
Suvo CHATTERJEE, *National Institutes of Health*
Shrabanti CHOWDHURY, *Icahn School of Medicine at Mount Sinai*
Sapatarshi CHATTERJEE, *Eli Lilly & Company*
Ali Rahnavard, Stephanie HICKS, *George Washington University, Johns Hopkins University*

The performance of computational methods and software to identify differentially expressed genes in single-cell RNA-sequencing (scRNA-seq) has been shown to be influenced by several factors, including the choice of the normalization method used and the choice of the experimental platform (or library preparation protocol) to profile gene expression in individual cells. Currently, it is up to the practitioner to choose the most appropriate differential expression (DE) method out of over 100 DE tools available to date, each relying on their own assumptions to model scRNA-seq data. Here, we propose to use generalized linear models with the Tweedie distribution that can flexibly capture a large dynamic range of observed scRNA-seq data across experimental platforms induced by heavy tails, sparsity, or different count distributions to model the technological variability in scRNA-seq expression profiles. We also propose a zero-inflated Tweedie model that allows zero probability mass to exceed a traditional Tweedie distribution to model zero-inflated scRNA-seq data with excessive zero counts. Using both synthetic and published plate- and droplet-based scRNA-seq datasets, we performed a systematic benchmark evaluation of more than 10 representative DE methods and demonstrate that our method (Tweedieverse) outperforms the state-of-the-art DE approaches across experimental platforms in terms of statistical power and false discovery rate control. Our open-source software (R package) is available at https://github.com/himelmallick/Tweedieverse.

118. Robust Variable Selection Criteria for the Penalized Regression

[IS 34, (page 16)]

Abhijit MANDAL, *Department of Mathematical Sciences, University of Texas at El Paso*
Samiran GHOSH, *Wayne State University*

In this talk, I will present a penalized robust variable selection procedure using a divergence based M-estimator. It produces robust estimates of the regression parameters and simultaneously selects the important explanatory variables. I will discuss on the asymptotic distribution and the influence function of the regression coefficients. The widely used model selection procedures based on the Mallows's C_p statistic and Akaike information criterion (AIC) often show very poor performance in the presence of heavy-tailed error or outliers. For this purpose, I will introduce robust versions of these information criteria based on our proposed method. The simulation studies show that the robust variable selection technique outperforms the classical likelihood-based techniques in the presence of outliers. The performance of the proposed method will also be explored through the real data analysis.

119. A Hierarchical Bayes Unit-Level Small Area Estimation Model for Normal Mixture Populations

[IS 13, (page 8)]

Abhyuday MANDAL, *University of Georgia*
Gauri S DATTA, *University of Georgia*
Adrijo CHAKRABORTY, *U.S. Food and Drug Administration*
Shuchi GOYAL, *University of California, Los Angeles*

National statistical agencies are regularly required to produce estimates about various subpopulations, formed by demographic and/or geographic classifications, based on a limited number of samples. Traditional direct estimates computed using only sampled data from individual subpopulations are usually unreliable due to small sample sizes. Subpopulations with small samples are termed small areas or small domains. To improve on the less reliable direct estimates, model-based estimates, which borrow information from suitable auxiliary variables, have been extensively proposed in the literature. However, standard model-based estimates rely on the normality assumptions of the error terms. In this research we propose a hierarchical Bayesian (HB) method for the unit-level nested error regression model based on a normal mixture for the unit-level error distribution. Our method proposed here is applicable to model cases with unit-level error outliers as well as cases where each small area population is comprised of two subgroups, neither of which can be treated as an outlier. Our proposed method is more robust than the normality based standard HB method (Datta and Ghosh 1991) to handle outliers or multiple subgroups in the population. Our proposal assumes two subgroups and the two-component mixture model that has been recently proposed by Chakraborty et al. (2018) to address outliers. To implement our proposal we use a uniform prior for the regression parameters, random effects variance parameter, and the mixing proportion, and we use a partially proper non-informative prior distribution for the two unit-level error variance components in the mixture. We apply our method to two examples to predict summary characteristics of farm products at the small area level. One of the examples is prediction of twelve county-level crop areas cultivated for corn in some Iowa counties. The other example involves total cash associated in farm operations in twenty-seven farming regions in Australia. We compare predictions of small area characteristics based on the proposed method with those obtained by applying the Datta and Ghosh (1991) and the Chakraborty et al. (2018) HB methods. Our simulation study comparing these three Bayesian methods, when the unit-level error distribution is normal, or t, or two-component normal mixture, showed the superiority of our proposed method, measured by prediction mean squared error, coverage probabilities and lengths of credible intervals for the small area means. (Joint research with Gauri Sankar Datta, Adrijo Chakraborty and Shuchi Goyal.)

120. On recurrent-event win ratio

[IS 24, (page 12)]

Lu MAO, *Biostatistics and Medical Informatics,University of Wisconsin-Madison*
KyungMann KIM, *University of Wisconsin-Madison*
Yi LI, *University of Wisconsin-Madison*

The win ratio approach proposed by Pocock et al. (2012) has become a popular tool for analyzing composite endpoints of death and non-fatal events like hospitalization. A limitation of the standard version of the win ratio is that it draws only on the first occurrence of the non-fatal event. For greater statistical efficiency and fuller characterization of patient experience, we construct and compare different win ratio variants that capture recurrent non-fatal events. We pay special attention to a variant called last-event-assisted win ratio (LWR), which compares two patients on the cumulative frequency of the non-fatal event, with ties broken by time to its last episode. It is shown that LWR uses more information than the standard win ratio does and reduces to the latter when the non-fatal event occurs at most once. We further prove that LWR rejects the null hypothesis with large probability if the treatment does stochastically delay all events. Simulations under realistic settings show that LWR consistently outperforms the standard win ratio and other competitors in statistical power. Analysis of a real cardiovascular trial provides further evidence for the practical advantages of LWR.

121. Completion and Classification of Partially Observed Curves with Application to Classification of Bovid Teeth

[IS 39, (page 18)]

Gregory MATTHEWS, *Mathematics and Statistics,Loyola University Chicago*
Ofer HAREL,
Karthik BHARATH,
Sebastian KURTEK,
Juliet BROPHY,

Statistical shape analysis of curves is well-developed when curves are fully observed. This work considers partially observed curves and develops methods for curve completion or imputation by leveraging tools from the statistical analysis of shape of fully observed curves, which enables sensible curve completions. On a dataset containing partially observed bovid teeth arising from a biological anthropology application, the method is implemented and classification of the completed teeth is carried out based on a shape distance on the set of curves.

122 . Causal Inference with the Instrumental Variable Approach and Bayesian Nonparametric Machine Learning

[IS 15, (page 8)]

Robert MCCULLOCH, *School of Mathematical and Statistical Sciences,Arizona State*
Purushottam LAUD, *Medical College of Wisconsin*
Brent LOGAN, *Medical College of Wisconsin*
Rodney SPARAPANI, *Medical College of Wisconsin*

We provide a new flexible framework for inference with the instrumental variable model. Rather than using linear specifications, functions characterizing the effects of instruments and other explanatory variables are estimated using machine learning via Bayesian Additive Regression Trees (BART). Error terms and their distribution are inferred using Dirichlet Process mixtures. Simulated and real examples show that when the true functions are linear, little is lost. But when nonlinearities are present, dramatic improvements are obtained with virtually no manual tuning.

123. Random Forests: Why They Work and Why That's a Problem

[IS 59, (page 27)]

Lucas MENTCH, *Department of Statistics,University of Pittsburgh*
Siyu ZHOU, *University of Pittsburgh*

Random forests remain among the most popular off-the-shelf supervised machine learning tools with a well-established track record of predictive accuracy in both regression and classification settings. Despite their empirical success, a full and satisfying explanation for their success has yet to be put forth. In this talk, we will show that the additional randomness injected into individual trees serves as a form of implicit regularization, making random forests an ideal model in low signal-to-noise ratio (SNR) settings. From a model-complexity perspective, this means that the mtry parameter in random forests serves much the same purpose as the shrinkage penalty in explicit regularization procedures like the lasso. Realizing this, we demonstrate that alternative forms of randomness can provide similarly beneficial stabilization. In particular, we show that augmenting the feature space with additional features consisting of only random noise can substantially improve the predictive accuracy of the model. This surprising fact has been largely overlooked in the statistics community, but has crucial implications for thinking about how to objectively define and measure variable importance. Numerous demonstrations on both real and synthetic data are provided.

124 . Computer model emulation for high dimensional functional output from satellite remote sensing

[IS 34, (page 16)]

Anirban MONDAL, *Department of Mathematics, Applied Mathematics, and Statistics,Case Western Reserve University*
Pulong MA, *SAMSI, Duke University*
Jonathan HOBBS, *Jet Propulsion Laboratory*
Emily KANG, *University of Cincinnati*
Konomi, BLEDAR, *University of Cincinnati*

NASA's Orbiting Carbon Observatory-2 (OCO-2) collects tens of thousands of observations of reflected sunlight daily, and the mission's retrieval algorithm processes these indirect measurements into estimates of atmospheric CO2 and other states. The physical forward model describing the mathematical relationship between the atmospheric state and the observed radiances is developed by the OCO-2 science team in the form of an expensive computer simulation. The multiple runs of this expensive computer simulation model make the retrieval algorithm computationally very expensive. Here we focus on the emulator approach where a statistical representation of the forward model is built based on some simulation runs of the forward model. Once the emulator is built it can be used to predict the observed radiance output for a given set of atmospheric state vectors almost instantaneously. We use functional principal component analysis to reduce the dimension of the functional radiance output and an active subspace approach to reduce the dimension input state vector. The nearest neighbor Gaussian process is used to fit the emulator and its performance is compared with a physical surrogate model.

125 . Adaptive Methods for Time-Modulated Variable Stars

[IS 22, (page 11)]

Giovanni MOTTA, *Statistics,Texas A&M, Department of Statistics*

In this talk we focus on Long Period Variable (LPV) and Blazhko stars, both characterized by slowly time-varying (or simply time-modulated) parameters: mean, amplitude, period and phase. Miras are a typical example of LPV stars, with an average mean period ranging from 100 to 1,000 days and large amplitudes of light variation of more than 2.5 magnitudes visually and more than 1 magnitude in the 5 infrared wavelengths. The period of these stars is a very useful indicator of their size and luminosity as well as their age, mode of pulsation and their overall evolution. Previous research has revealed some important correlations between the period and other parameters such as amplitude, mass loss and IR excess due to dust surrounding the star. The magnitude of LPV exhibits a (possibly quadratic) timevarying mean, as well as time-varying amplitude and period. The Blazhko effect, which is sometimes called long-period modulation, is a variation in period, amplitude or phase in RR Lyrae type variable stars. The amplitude-modulated pulsation of RR Lyrae stars has a strong periodic component with an often observed variation on a longer time scale. The amplitude variation is accompanied by phase changes of the same period. The modulation period can be anywhere between 10 and 700 days, without any correlation with the fundamental period. The Blazhko effect is a periodic amplitude and/or phase modulation shown by some 20-30% of the galactic RRab stars. Our goals are modeling and forecasting these light curves. In our approach we allow for a smooth time-varying trend, as well as for smooth time-varying coefficients describing the local (in time) amplitudes of the cosine and sine waves. Our approach is flexible because it avoids assumptions about the functional form of trend and amplitudes. More precisely, we propose a semi-parametric model where only part of the model is time-varying. The estimation of our time-varying curves translates into the estimation of time-invariant parameters that can be performed by ordinary least-squares, with the following two advantages: modeling and forecasting can be implemented in a parametric fashion, and we are able to cope with missing observations.

126 . Dependent Mixtures: Modeling Cell Lineage

[IS 58, (page 26)]

Peter MUELLER, *Statistics & Data Sc,UT Austin*
Carlos Pagani ZANINI, *UFRJ*
Giorgio PAULON, *UT Austin*

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.

127. Improved Nonparametric Empirical Bayes Estimation using Transfer Learning

[IS 9, (page 6)]

Gourab MUKHERJEE, *Department of Data Sciences & Operations,University of Southern California*

We consider the problem of estimating a multivariate normal mean in the presence of possibly useful auxiliary variables. The traditional nonparametric empirical Bayes (NEB) framework provides an elegant interface to pool information across dimensions and facilitates the construction of effective shrinkage estimators. Such estimators can be further improved by incorporating pertinent information from the auxiliary variables. However, detecting and assimilating possibly useful information from auxiliary variables to shrinkage estimators is difficult. Here, we develop a new methodology that can transfer useful information from multiple auxiliary variables and yield improved Tweedie-type NEB estimators. Our method uses convex optimization to directly estimate the gradient of the log-density through an embedding in the reproducing kernel Hilbert space induced by the Stein's discrepancy metric. We establish asymptotic optimality of the resultant estimator. We precisely tabulate the improvements in the estimation error as well as the deterioration in the learning rate as we inspect an increasing number of auxiliary variables. We demonstrate the competitive optimality of our method over existing NEB approaches through simulation experiments and in real data settings.

128. Optimal offline changepoint estimation in network sequences

[IS 31, (page 15)]

Soumendu Sundar MUKHERJEE, *Interdisciplinary Statistical Research Unit (ISRU),Indian Statistical Institute, Kolkata*
Sharmodeep BHATTACHARYYA, *Oregon State University*
Shirshendu CHATTERJEE, *City University of New York*
Trisha DAWN, *Indian Statistical Institute, Kolkata*
Tamojit SADHUKHAN, *Indian Statistical Institute, Kolkata*

We will present ongoing work on offline changepoint estimation in network-valued time series. The networks involved are inhomogeneous ErdsRnyi random graphs. For the most part, we will assume them to be independent across time, although we will present some results for certain types of time-dependent models as well. We will also consider submodels with structural changes, such as changes in community structure. We will present CUSUM-type estimators which are optimal in an appropriate minimax sense.

129 . Global testing for dependent Bernoullis

[IS 17, (page 9)]

Sumit MUKHERJEE, *Sumit Mukherjee,Columbia University*
Nabarun DEB, *Columbia University*
Rajarshi MUKHERJEE, *Harvard University*
Ming YUAN, *Columbia University*

Suppose $(X_1, \ldots, X_n)$ are independent Bernoulli random variables with $\mathbb{E}(X_i) = p_i$, and we want to test the global null hypothesis that $p_i = \frac{1}{2}$ for all i, versus the alternative that there is a sparse set of size s on which $p_i \geq \frac{1}{2} + A$. The detection boundary of this test in terms of (s, A) is well understood, both in the case when the signal is arbitrary, and when the signal is present in a segment.

We study the above questions when the Bernoullis are dependent, and the dependence is modeled by a graphical model (Ising model). In this case, contrary to what typically happens, dependence can allow detection of smaller signals than the independent case. This phenomenon happens over a wide range of graphs, for both arbitrary signals and segment signals.

130. Evaluation of Logrank and Max-Combo test in Immuno-Oncology Trials A Retrospective Analysis in Patients Treated with Anti-PD1/PD-L1 Agents across Solid Tumors

[IS 44, (page 21)]

Pralay MUKHOPADHYAY, *Department of Biometrics,Otsuka America Pharmaceuticals Inc.*
Jiabu YE, *AstraZeneca*

The challenges of designing and analyzing of immune-oncology (IO) studies and the issue of non-proportional hazards (NPH) have now been well studied. It has been demonstrated that the Logrank test

(LRT) may be considerably under powered under NPH. Various alternative approaches have been proposed to overcome this limitation. This includes a range of option from sizing trials appropriately, accounting for NPH through larger sample size and longer duration of follow-up while using LRT for analysis. Alternatively using tests such as Fleming Harrington weighted LRTs (wLRT) or combination of wLRTs like the MC test. The debate is ongoing on use of wLRTs as a method for primary analysis owing to its over weighting parts of the Kaplan-Meier (KM) curve versus treating all events as equal. There have also been examples cited of weighted tests not controlling type I error under the strong null and may have tendency to declare statistical significance when the clinical relevance of data may be questionable.

In this session, we retrospectively compare the LRT with MC across 57 IO studies in approximately 33562 patients treated with different anti-PD1/PD-L1 therapies in various solid tumors. A total of 139 statistical comparisons are made between the experimental and control arm, on the all-comers or PD-L1+ subgroups. The KM curves were digitized from available publications to obtain the event and censoring information for analysis. We focus our review on the discordant cases where either the MC was statistically significant while the LRT was not or vice versa. We look at the clinical relevance of these results and share observations around the use of LRT versus MC in practice and its potential implication in IO drug development. We also explore how to best describe treatment effect in many of these cases where severe NPH was observed.

131. Distribution free estimation of optimal order quantity for a newsboy

[IS 4, (page 4)]

Sujay MUKHOTI, *Operations Management and Quantitative Techniques,Indian Institute of Management Indore*

Classical newsvendor's problem deals with optimal order quantity determination at the beginning of the sales period for perishable goods. The optimal order quantity is so decided that it balances the losses due to shortage and excess. In this work, we propose a generalization of the newsvendor problem with different degrees of sensitivity towards shortage and excess, as opposed to the classical newsvendor problem, which treats these two with equal importance. We propose two weight functions for the shortage and excess costs. We discuss both the cases of equal and unequal weight functions. In case of equal weight functions, we present the case of power-type weights. Further, we have proposed a non-parametric estimation of the optimal order quantity, when the demand distribution is unknown. We provide a brief discussion on the effect of misspecification on the estimators.

132. Visual Analytics, Big Data, and Drug Safety with a Focus on Text Mining and Natural Language Processing

[IS 7, (page 5)]

Melvin MUNSAKA, *Statistical Sciences,AbbVie*
Kefei ZHOU, *Bristol Myers Squibb*
Krishan SINGH, *GlaxoSmithKline (Retired)*

The availability of medical and healthcare data from electronic resources coupled with readily available high speed computing resources has opened new frontiers and opportunities for the assessment of drug safety. Of note, these real-world data sources fall under the realm of big data in terms of volume, veracity, velocity, and variety. Visualizing these data can be useful to various stakeholders. As an example, for clinicians and quantitative scientists, comprehension of patients medical history and concurrent medications are important in characterizing drug safety profiles. Visual analytics facilitates for blending of data visualization and statistical and data mining techniques to create visual forms that help researchers make sense of safety data with emphasis on how to complement computation and visualization. This is important in enabling maximum information gain from visual analytics and for more effective and informative visualization. This discussion will focus on visual analytics of real-world data sources with a focus on web and social media sources. We will also discuss open-source tools that can be leveraged for visual analytics of these data.

133. Bayesian Multiple Quantile Regression Using a Score Likelihood

[IS 40, (page 18)]

Naveen Naidu NARISETTY, *Statistics,University of Illinois at Urbana-Champaign*
Teng WU, *University of Illinois at Urbana-Champaign*

We study the use of a score based working likelihood function for quantile regression which can perform inference for multiple conditional quantiles of an arbitrary number. The proposed likelihood can be used in a Bayesian framework leading to valid frequentist inference, whereas the commonly used asymmetric Laplace working likelihood leads to invalid interval estimations and requires further correction. We present a novel adaptive importance sampling algorithm to compute important posterior summaries such as the posterior mean and the covariance matrix. Our proposed approach makes it feasible to perform valid inference for parameters such as the slope differences at different quantile levels, which is either not possible or cumbersome using existing Bayesian approaches. Empirical results demonstrate that the proposed likelihood has good estimation and inferential properties and the proposed algorithm is more efficient than competitors.

134. Drug Development in the 21st Century Need for Innovation in Statistical Thinking

[Plenary Lecture 2, (page 13)]

Kannan NATARAJAN, *Global Head of Biometrics and Data Management,Pfizer*

In the 21st century, drug discovery and development face a dramatically different environment and challenges than those of just a few years ago. Significant advances in science are made in genetic, molecular, and cellular levels provide greater opportunity for more effective therapies. This has occurred in search of better and more targeted medicines that reach patients with unmet medical based on disease histopathology. Traditional statistical principles are often not ideal to establish the benefit risk of a new class of medical products under this changing landscape. Therefore, new statistical thinking is essential to help sponsors, decision-makers, and regulators for making good decision about medical products and market authorization as quickly and efficiently as possible in the presence of uncertainty. This talk will reflect some of the recent statistical innovations that have increased the efficiency of medical product development and facilitate robust regulatory decision making.

The talk will begin with some recent innovations in the rare disease area. With a significant unmet medical need in some of the diseases and relatively fewer patients, it is unethical and often not feasible to conduct large trials with either placebo or suboptimal active controls. The talk will focus on the use of historical control information, natural disease history records, and Real-World Data (RWD) to enrich the design and analysis of clinical trial in rare disease. The second half of the talk will cover some of the innovative approaches used in the recent Pfizer/BioNTech COVID-19 vaccine trial. The talk will include 1) Bayesian approach used for interim monitoring and reporting of vaccine efficacy, and 2) an AI/Machine Learning platform to assess complex data discrepancies, reconciling data and managing data queries. The talk will conclude with some remarks regarding the future direction of clinical trials (e.g., use of decentralized trials) and recent regulatory developments.

135. Empirical Bayes and the false discovery rate, revisited

[Special Invited Session 1, (page 7)]

Michael NEWTON, *Department of Statistics and Department of Biostatistics and Medical Informatics,University of Wisconsin-Madison*

Large-scale hypothesis testing and prioritization problems occur in many contemporary statistical applications. Well developed and widely studied methods and computational tools are available to report the most interesting testing units subject to control on the rate of false discoveries. Motivated by examples from single-cell RNA-sequencing, antibody profiling, and brain imaging, I will review recent work on empirical Bayesian approaches to this problem, noting in many cases that standard tools control false discovery rates but at the expense of deficiencies in other operating characteristics.

136. Causal Inference Under Interference in Dynamic Therapy Group Studies

[Special Invited Session 2, (page 7)]

Susan PADDOCK, *Statistics and Methodology,NORC at the University of Chicago*
Bing HAN, *RAND Corporation*
Lane BURGETTE, *RAND Corporation*

Group therapy is a common treatment modality for behavioral health conditions. Patients often enter and exit groups on an ongoing basis, leading to dynamic therapy groups. Examining the effect of high versus low session attendance on patient outcomes is of interest. However, there are several challenges to identifying causal effects in this setting, including the lack of randomization, interference among patients, and the interrelatedness of patient participation. Dynamic therapy groups motivate a unique causal inference scenario, as the treatment statuses are completely defined by the patient attendance record for the therapy session, which is also the structure inducing interference. We adopt the Rubin Causal Model framework to define the causal effect of high versus low session attendance of group therapy at both the individual patient and peer levels. We propose a strategy to identify individual, peer, and total effects of high attendance versus low attendance on patient outcomes by the prognostic score stratification. We examine performance of our approach via simulation, apply it to data from a group cognitive behavioral therapy trial for reducing depressive symptoms among patients in a substance use disorders treatment setting, and discuss the strengths and limitations of this approach.

137 . Median-Unbiasedness in Finite Population Survey Sampling

[IS 10, (page 6)]

Jennifer PAJDA-DE LA O, *Mathematics, Statistics, and Computer Science,University of Illinois at Chicago*
A. S. HEDAYAT, *University of Illinois at Chicago*

There are no unbiased estimators of the minimum, maximum, or median for finite population sampling under any sampling design except census. However, it is possible to find estimators that are median-unbiased. A simple random sampling design has the median-unbiasedness property. By deleting samples from a simple random sample and imposing a uniform probability distribution on the remaining samples, the sample median is a median-unbiased estimator, provided that the support meets a minimum threshold. It is possible to construct other types of sampling designs that do not have a minimum threshold requirement where the sample median is a median-unbiased estimator.

138. The Role of Natural History Data in Rare Disease Clinical Development

[IS 19, (page 10)]

Jeff PALMER, *Early Clinical Development Statistics,Pfizer, Inc.*

One of the primary challenges in rare disease drug development continues to be the small numbers of patients available for participation in interventional clinical trials. Researchers have made progress in adopting novel statistical methods for dealing with this challenge such as the use of composite endpoints, Bayesian trial designs, and incorporation of patient-level data from various natural history data sources. Natural history data is critical in rare disease drug development as it can help facilitate understanding of disease progression, aid in the selection of clinically meaningful and sensitive endpoints, and in some cases, be used to augment the observed placebo response in an interventional trial. In this presentation we will review how natural history data is being used to support the rare disease portfolio at Pfizer.

139. Statistical optimality and stability of tangent transform algorithms

[IS 64, (page 29)]

Debdeep PATI, *Statistics,Texas A&M University*
Anirban BHATTACHARYA, *Texas A&M University*
Indrajit GHOSH, *Texas A&M University*

A systematic approach to finding variational approximation in an otherwise intractable non-conjugate model is to exploit the general principle of convex duality by minorizing the marginal likelihood using a convex minorant that renders the problem tractable. While such approaches are popular in the context of variational inference in non-conjugate Bayesian models, theoretical guarantees on statistical optimality and algorithmic convergence are lacking. Focusing on logistic regression models, we provide mild conditions on the data generating process to derive non-asymptotic upper bounds to the risk incurred by the variational optima. We demonstrate that these assumptions can be completely relaxed if one considers a slight variation of the algorithm by raising the likelihood to a fractional power. Next, we utilize the theory of dynamical systems to provide convergence guarantees for such algorithms in logistic and multinomial logit regression. In particular, we establish local asymptotic stability of the algorithm without any assumptions on the data-generating process. We explore several special cases involving an orthogonal and a semi-orthogonal design under which a global convergence rate is obtained. The theory is further illustrated using several numerical studies.

140 . Least Squares Estimation of a Monotone Quasiconvex Regression Function

[IS 5, (page 5)]

Rohit PATRA, *Department of Statistics,University of Florida*
Somabha MUKHERJEE, *University of Pennsylvania*
Andrew L JOHNSON, *Amazon*
Hiroshi MORITA, *Osaka University*

We develop a new approach for the estimation of a multivariate function based on the economic axioms of monotonicity and quasiconvexity. We prove the existence of the nonparametric least squares estimator (LSE) for a monotone and quasiconvex function and provide two characterizations for it. One of these characterizations is useful from the theoretical point of view, while the other helps in the computation of the estimator. We show that the LSE is almost surely unique and is the solution to a mixed-integer quadratic optimization problem. We prove consistency and find finite sample risk bounds for the LSE under both fixed lattice and random design settings for the covariates. We illustrate the superior performance of the LSE against existing estimators via simulation. Finally, we use the LSE to estimate the production function for the Japanese plywood industry and the cost function for hospitals across the US.

141. New Approaches for Testing Non-inferiority for Three-arm Trials with Poisson Distributed Outcomes

[IS 27, (page 13)]

Erina PAUL, *Biostatistics and Research Decision Sciences,Merck & Co., Inc.*
Samiran GHOSH,
Shrabanti CHOWDHURY,
Ram TIWARI,

With the availability of limited resources, innovation for improved statistical method for the design and analysis of randomized controlled trials (RCTs) are of paramount importance for newer and better treatment discovery for any therapeutic area. Although clinical efficacy is almost always the primary evaluating criteria to measure any beneficial effect of a treatment, there are several important other factors (e.g., side effects, cost burden, less debilitating, less intensive etc.), which can permit some less efficacious treatment options favorable to a subgroup of patients. This leads to non-inferiority (NI) testing. The objective of NI trial is to show that an experimental treatment is not worse than an active reference treatment by more than a pre-specified margin. Traditional NI trials do not include a placebo arm for ethical reason, however, this necessitates stringent and unverifiable assumptions. On the other hand, three-arm NI trials consisting of placebo, reference, and experimental treatment, can simultaneously test the superiority of the reference over placebo and NI of experimental treatment over the reference. In this paper, we consider both Frequentist and Bayesian procedures for testing NI in the three-arm trial with Poisson distributed count outcome. RCTs with count data as the primary outcome are quite common in various disease areas such as lesion count in cancer trials, relapses in multiple sclerosis, dermatology, neurology, cardiovascular research, adverse event count, etc. We first propose an improved Frequentist approach, which is then followed by a Bayesian version. Bayesian methods has natural advantage in any active-control trials, including NI trial when substantial historical information is available for placebo and established reference treatment. In addition, we discuss sample size calculation and draw an interesting connection between the two paradigms.

142 . Modeling continuous-time networks of relational events

[IS 36, (page 17)]

Subhadeep PAUL, *Department of Statistics,The Ohio State University*
Kevin XU, *University of Toledo*
Makan ARASTUIE, *University of Toledo*

In many application settings involving networks, such as messages between users of an on-line social network or transactions between traders in financial markets, the observed data consist of timestamped relational events, which form a continuous-time network. We propose the Community Hawkes Independent Pairs (CHIP) generative model for such networks. We show that applying spectral clustering to an aggregated adjacency matrix constructed from the CHIP model provides consistent community detection for a growing number of nodes and time duration. We also develop consistent and computationally efficient estimators for the model parameters. We demonstrate that our proposed CHIP model and estimation procedure scales to large networks with tens of thousands of nodes and provides superior fits than existing continuous-time network models on several real networks.

143. Fitting Proportional Odds Model with Missing Responses When the Missing Data Are Nonignorable

[IS 49, (page 22)]

Vivek PRADHAN, *Statistics, ECD I&I,Pfizer Inc*

In clinical trial and many other applications proportional odds model are fitted to model ordered cat-

egorical data. In real world application, especially in clinical trial data, presences of missing data are inevitable. Missing values in the data occur due to different missing mechanisms, such as missing completely at random (MCAR) and missing at random (MAR). In the regression set up with missing data, most of the focuses are on missing covariates; however, very little attention has been paid to missing responses. If the responses are missing, and the missingness depends on the responses itself, then it is called nonignorable missing. In our work, we will focus on how to handle missing responses in fitting proportional odds model when the missing data are nonignorable. Following Ibrahim and Lipsitz (Biometrics 52: 1071-1078, 1996) we will propose an EM algorithm to fit the model. We will investigate the bias correction of the estimates of the regression coefficients. All of these novel methods will be demonstrated using real world data.

144. Bayesian Design of Pediatric Clinical Trials with Prospective Incorporation of Data from Adult Trials

[IS 21, (page 11)]

Matthew PSIODA, *Department of Biostatistics,Department of Biostatistics*

Two common features of pediatric drug development are that (1) pediatric populations are often difficult to enroll in trials and (2) data from adult trials is often available before pediatric trials begin. Motivated by these challenges, we propose a sequential monitoring methodology where information from adult trials is prospectively incorporated using a skeptical power prior. Information borrowing from adults is limited based on the pediatric trial sample size so as to not overwhelm the pediatric data until those data are reasonable mature. In particular, we address the challenging issue of information borrowing from large information sources (e.g., adult trials) in the analysis of accumulating data that contribute little information a common characteristic for pediatric trials. This is achieved through adaptive information borrowing using a power prior with discounting parameter computed using the prior predictive distribution for the pediatric data given the adult data, and principles for rational decision making that are critical when the information in the pediatric data is not substantial, making so called prior-data conflict difficult to discern.

145. Randomization tests of causal effects under general interference

[IS 56, (page 26)]

David PUELZ, *Booth School of Business,University of Chicago*

Interference exists when a units outcome depends on another units treatment assignment. For example, intensive policing on one street could have a spillover effect on neighboring streets. Classical randomization tests typically break down in this setting because many null hypotheses of interest are no longer sharp under interference. A promising alternative is to instead construct a conditional randomization test on a subset of units and assignments for which a given null hypothesis is sharp. Finding these subsets is challenging, however, and existing methods are limited to special cases or have limited power. In this paper, we propose valid and easy-to- implement randomization tests for a general class of null hypotheses under arbitrary interference between units. Our key idea is to represent the hypothesis of interest as a bipartite graph between units and assignments, and to find an appropriate biclique of this graph. Importantly, the null hypothesis is sharp within this biclique, enabling conditional randomization-based tests. We also connect the size of the biclique to statistical power. Moreover, we can apply off-the-shelf graph clustering methods to find such bicliques efficiently and at scale. We illustrate our approach in settings with clustered interference and show advantages over methods designed specifically for that setting. We then apply our method to a large-scale policing experiment in Medelln, Colombia, where interference has a spatial structure.

146. Convergence complexity analysis of MCMC algorithm

[IS 15, (page 8)]

Qian QIN, *School of Statistics,University of Minnesota*
James P. HOBERT, *University of Florida*

Convergence complexity analysis is the study of how the convergence behavior of Monte Carlo Markov chains scales with sample size, n, and/or number of covariates, p. Methods based on Wasserstein distances and random mappings can produce bounds that are robust to increasing dimension. The Wasserstein-based bounds are used to develop strong convergence complexity results for MCMC algorithms used in Bayesian probit regression and random effects models in the challenging asymptotic regime where n and p are both large.

147. Recent advances in defining estimands and imputation of missing data in clinical trials

[IS 1, (page 3)]

Yongming QU, *Department of Data and Analytics,Eli Lilly and Company*

ICH E9(R1) Addendum, released in 2019, discusses the process of defining estimands, especially on strategies for handling intercurrent events and sensitivity analyses. In this presentation, we will discuss various hypothetical strategies for handling intercurrent events as well as the corresponding methods of imputing missing values. Specifically, to handle intercurrent events, we will discuss the control direct hypothetical, partial treatment hypothetical, and no treatment hypothetical strategies. To handle missing data, we will discuss commonly used multiple imputation methods including imputation under the assumption of missing at random and under special patterns (retrieved dropout imputation, reference-based imputation, return-to-baseline imputation, and imputation with sensitivity parameter).

148. Omics Community Detection using Multi-resolution Clustering

[IS 18, (page 10)]

Ali RAHNAVARD, *Biostatistics and Bioinformatics,George Washington University*
Himel MALLICK, *Merck & Co., Inc.*
Suvo CHATTERJEE, *National Institutes of Health*
Bahar SAYOLDIN, *George Mason University*
Keith A. CRANDALL, *George Washington University*

Finding biologically interpretable and clinically actionable communities in heterogeneous omics data is a necessary first step towards deriving mechanistic insights into complex biological phenomena. Here we present a novel clustering approach, omeClust, for community detection in omics profiles by simultaneously incorporating similarities between measurements and the overall complex structure of the data. We show that omeClust outperforms published methods in inferring the true community structure as measured by both sensitivity and misclassification rate on simulated datasets. We validated omeClust in diverse, multiple omics datasets, revealing new communities and functionally related groups in microbial strains, cell line gene expression patterns, and fetal genomic variation. We further derived enrichment scores attributable to putatively meaningful biological factors in these datasets that can serve as a hypothesis generator and facilitate new sets of testable hypotheses. omeClust is open-source software and the implementation is available online at http://github.com/omicsEye/omeClust.

149. Techniques and methods for data visualization in clinical trials illustrated with examples

[IS 8, (page 6)]

Mallikarjuna RETTIGANTI, *Neuroscience,Eli Lilly and Company*
Bochao JIA, *Eli Lilly and Company*
Eric WOLF, *Eli Lilly and Company*
Jakub JEDYNAK, *Eli Lilly and Company*

Data visualization can be a very powerful tool to convey data narratives that cannot be effectively communicated using tables or text. In recent years, powerful visualization methods and techniques such as motion graphics have been used to highlight trends and patterns in clinical trial data that cannot be otherwise seen in static graphs or grouped data. In this talk, we will use real life examples from across various therapeutic areas to give a flavor of visualizations and animations currently employed in the pharmaceutical industry. In migraine trials, we use animated raindrop plots to show trends in patient level data over time. Further, we utilize animated bar plots to demonstrate movement between migraine headache categories over time within each patient. We also illustrate the use of spider animation plots to visualize the speed of onset of efficacy and magnitude of improvement for multiple endpoints as applicable to atopic dermatitis clinical trials. We will also share the different tools/methods used to produce these visualizations discussed. When used effectively, new trends in data visualizations can vastly improve how data are represented in clinical trials in a way that relates to the patients, health care practitioners and other important stakeholders.

150. Bayesian time-aligned factor analysis of paired multivariate time series

[IS 4, (page 4)]

Arkaprava ROY, *University of Florida*
David DUNSON, *Duke University*
Jana SCHAICH-BORG, *Duke University*

Many modern data sets require inference methods that can estimate the shared and individual-specific components of variability in collections of matrices that change over time. Promising methods have been developed to analyze these types of data in static cases, but only a few approaches are available for dynamic settings. To address this gap, we consider novel models and inference methods for pairs of matrices in which the columns correspond to multivariate observations at different time points. In order to characterize common and individual features, we propose a Bayesian dynamic factor modeling framework called Time Aligned Common and Individual Factor Analysis (TACIFA) that includes uncertainty in time alignment through an unknown warping function. We provide theoretical support for the proposed model, showing identifiability and posterior concentration. The structure enables efficient computation through a Hamiltonian Monte Carlo (HMC) algorithm. We show excellent performance in simulations, and illustrate the method through application to a social synchrony experiment.

151. Use of External Controls in Oncology Trials

[IS 30, (page 14)]

Pourab ROY, *Office of Biostatistics,FDA*

Randomized clinical trials (RCT) remain the gold standard for the identification of treatment effect, however, disease or population characteristics may require a non-randomized study design. In such cases, an external control arm may be utilized for estimating comparative treatment effect. In this presentation we will go over the different considerations while using external control data as well as a short case study to demonstrate its applicability.

152. MCMC algorithms for Bayesian generalized linear mixed models

[IS 35, (page 17)]

Vivekananda ROY, *Department of Statistics,Iowa State University*

The likelihood function in generalized linear mixed models (GLMMs) is available only as a high dimensional integral, and thus the resulting posterior densities in Bayesian GLMMs are intractable. Generally, Markov chain Monte Carlo algorithms are used to explore these posterior densities. In this talk, we will consider two popular GLMMs, namely the probit linear mixed models and the logistic linear mixed models. We will present some block Gibbs (BG) samplers constructed using the data augmentation techniques for both of these GLMMs. We will also provide conditions guaranteeing geometric ergodicity of these BG Markov chains. This theoretical result has important practical implications as it justifies the use of asymptotically valid Monte Carlo standard errors for Markov chain based estimates of posterior quantities. Finally, we will present some numerical examples to demonstrate superior performance of the BG samplers over the full Gibbs samplers.

153. Kernel Distance Covariance Approach for Testing Association in Longitudinal Studies

[IS 12, (page 7)]

Pratyaydipta RUDRA, *Statistics,Oklahoma State University*

Many gene mapping studies of complex traits have identified genes or variants that demonstrate pleiotropic effects and influence multiple distinct phenotypes. With the advent of next-generation sequencing technology, there has been substantial interest in identifying rare variants in genes that possess cross-phenotype effects. In the presence of such cross-phenotype effects, modeling both the phenotypes and rare variants collectively using multivariate models can achieve higher statistical power compared to univariate methods that either model each phenotype separately or perform separate tests for each variant. Several studies collect phenotypic data over time and using such longitudinal data can further increase the power to detect genetic associations. While rare-variant approaches often use kernel machine methods for testing cross-phenotype effects at a single time point, it is unclear how to use them for longitudinal data. We propose an extension of Gene Association with Multiple Traits (GAMuT) test, a method for cross-phenotype analysis of rare variants using a framework based on the kernel distance covariance. The approach allows for both binary and continuous phenotypes and can also adjust for covariates. Our simple adjustment to the GAMuT test allows it to handle longitudinal data and to gain power by exploiting temporal correlation. The approach is computationally efficient and applicable on a genome-wide scale due to the use of a closed-form test whose significance can be evaluated analytically. We use simulated data to demonstrate that our method has favorable power over competing approaches and also apply our approach to exome chip data from the Genetic Epidemiology Network of Arteriopathy.

154. Collaborative Design for Improved Causal Machine Learning on Big Observational Data

[IS 3, (page 4)]

Arman SABBAGHI, *Department of Statistics,Purdue University*
Yumin ZHANG, *Purdue University Department of Statistics*

A fundamental issue in causal machine learning for Big Observational Data is confounding due to covariate imbalances between treatment groups. This can be addressed by designing the data prior to analysis. Existing design methods, developed for traditional observational studies with single designers, can yield unsatisfactory designs with suboptimum covariate balance for Big Observational Data due to their inability to accommodate the massive dimensionality, heterogeneity, and volume of the Big Data. We propose a new framework for the distributed design of Big Observational Data amongst collaborative designers. Our framework first assigns subsets of the high-dimensional and heterogeneous covariates to multiple designers. The designers then summarize their covariates into lower-dimensional quantities, share their summaries with the others, and design the study in parallel based on their assigned covariates and the summaries they receive. The final design is selected by comparing balance measures for all covariates across the candidates. We perform simulation studies and analyze datasets from the 2016 Atlantic Causal Inference Conference Data Challenge to demonstrate the flexibility and power of our framework for constructing designs with good covariate balance from Big Observational Data.

155 . Trend filtering with sub-exponential noise and for exponential families

[IS 47, (page 22)]

Veeranjaneyulu SADHANALA, *Booth School of Business,University of Chicago*
Robert BASSETT, *Naval Postgraudate School*
James SHARPNACK, *University of California, Davis*
Dan MCDONALD, *University of British Columbia, Vancouver*

Trend filtering is a nonparametric regression method that adapts to local level of smoothness. We derive error bounds for the method over lattice graphs when the response is corrupted by sub-exponential noise. In a homoscedastic setting, these error bounds differ from the bounds in sub-Gaussian case by only a logarithmic factor. We also study trend filtering for natural exponential families and derive excess risk bounds.

156 . Multiple Change Point Detection in Reduced Rank High Dimensional Vector Autoregressive Models

[IS 11, (page 7)]

Abolfazl SAFIKHANI, *University of Florida,University of Florida*
Peiliang BAI, *University of Florida*
George MICHAILIDIS, *University of Florida*

In this talk, we discuss the problem of detecting and locating change points in high-dimensional Vector Autoregressive (VAR) models, whose transition matrices exhibit low rank plus sparse structure. We first address the problem of detecting a single change point using an exhaustive search algorithm and establish a finite sample error bound for its accuracy. Next, we extend the results to the case of multiple change points that can grow as a function of the sample size. Their detection is based on a two-step algorithm, wherein the first step, an exhaustive search for a candidate change point is employed for overlapping windows, and subsequently a backwards elimination procedure is used to screen out redundant candidates. The two-step strategy yields consistent estimates of the number and the locations of the change points. To reduce computation cost, we also investigate conditions under which a surrogate VAR model with a weakly sparse transition matrix can accurately estimate the change points and their locations for data generated by the original model. This work also addresses and resolves a number of novel technical challenges posed by the nature of the VAR models under consideration. The effectiveness of the proposed algorithms and methodology is illustrated on both synthetic and real data sets.

157. Test for Isotropy on a Sphere using Spherical Harmonic Coefficients

[IS 66, (page 30)]

Indranil SAHOO, *Statistical Sciences & Operations Research,Virginia Commonwealth University*
Joseph GUINNESS, *Cornell University*
Brian J. REICH, *North Carolina State University*

Analyses of geostatistical data are often based on the assumption that the spatial random field is isotropic. This assumption, if erroneous, can adversely affect model predictions and statistical inferences. Today, many applications consider global data, and hence, it is necessary to check the assumption of isotropy on a sphere. This study proposes a test for spatial isotropy on a sphere. The data are first projected onto the set of spherical harmonic functions. Under isotropy, the spherical harmonic coefficients are uncorrelated, but are correlated if the underlying fields are not isotropic. This motivates a test based on the sample correlation matrix of the spherical harmonic coefficients. In particular, we use the largest eigenvalue of this matrix as the test statistic. Extensive simulations are conducted to assess the Type-I errors of the test under different scenarios. Our method requires temporal replication in the data and, hence, is applicable to many data sets in the Earth sciences. We show how temporal correlation affects the test and provide a method for handling such correlation. We also gauge the power of the test as we move away from isotropy. The method is applied to near-surface air temperature data, which is part of the HadCM3 model output. Although we do not expect global temperature fields to be isotropic, we propose several anisotropic models, with increasing complexity, each of which has an isotropic process as a model component. Then, we apply the test to the isotropic component in a sequence of such models to determine how well the models capture the anisotropy in the fields.

158. Bayesian Semiparametric Longitudinal Functional Mixed Models with Locally Informative Predictors

[IS 41, (page 19)]

Abhra SARKAR, *Statistics and Data Sciences,The University of Texas at Austin*
Giorgio PAULON, *The University of Texas at Austin*
Peter MUELLER, *The University of Texas at Austin*

We present a flexible Bayesian semiparametric mixed model for longitudinal functional data analysis in the presence of potentially high-dimensional categorical covariates. Our proposed method allows the fixed effects components to vary between dependent random partitions of the covariate space at different time points. The mechanism not only allows different sets of covariates to be included in the model at different time points but also allows the selected predictors influences to vary flexibly over time. Smooth time-varying additive random effects are used to capture subject-specific heterogeneity. We establish posterior convergence guarantees for both function estimation and variable selection. We design a Markov chain Monte Carlo algorithm for posterior computation. We evaluate the methods empirical performances through synthetic experiments and demonstrate its practical utility through real-world applications.

159. Trading off Accuracy for Speedup: Multiplier Bootstraps for Subgraph Counts

[IS 31, (page 15)]

Purnamrita SARKAR, *Department of Statistics and Data Sciences,Asst. Prof.*
Qiaohui LIN, *Student, UT Austin*
Robert LUNDE, *Postdoctoral Scholar, UT Austin*

We propose a new class of multiplier bootstraps for count functionals. We consider bootstrap procedures with linear and quadratic weights. These correspond to the first and second-order terms of the Hoeffding decomposition of the bootstrapped statistic arising from the multiplier bootstrap, respectively. We show that the quadratic bootstrap procedure achieves higher-order correctness for appropriately sparse graphs. The linear bootstrap procedure requires fewer estimated network statistics, leading to improved accuracy over its higher-order correct counterpart in sparser regimes. To improve the computational properties of the linear bootstrap further, we consider fast sketching methods to conduct approximate subgraph counting and establish consistency of the resulting bootstrap procedure. We complement our theoretical results with a simulation study and real data analysis and verify that our procedure offers state-of-the-art performance for several functionals.

160. Bayesian network models for integrating genetics and metabolomics data

[IS 48, (page 22)]

Denise SCHOLTENS, *Department of Preventive Medicine - Biostatistics,Department of Preventive Medicine - Biostatistics*

Integration of genetics and metabolomics data demands careful accounting of complex dependencies, particularly when modeling familial omics data, for example, to study fetal programming of related maternal-offspring phenotypes. Efforts to find genetically determined metabotypes using classic GWAS approaches have proven useful for characterizing complex disease, but conclusions are often limited to a disjointed series variant-metabolite associations. We adapted Bayesian network models to integrate metabotypes with maternal-fetal genetic dependencies and metabolic profile correlations. Using data from the multiethnic Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study, we demonstrate that strategic specification of ordered dependencies, pre-filtering of candidate metabotypes, clustering of metabolites and conditional linear Gaussian methods clarify fetal programming of newborn adiposity related to maternal glycemia. Exploration of network growth over a range of penalty parameters, coupled with interactive plotting, facilitate interpretation of network edges. These methods are broadly applicable to integration of diverse omics data for related individuals.

161. A Bayesian Joint Model for Clustered Agreement Data

[IS 29, (page 14)]

Ananda SEN, *Department of Biostatistics, University of Michigan*
Wen YE, *University of Michigan*
Pin LI, *Henry Ford Health System*

In medical imaging, inter and intra-rater agreement measures provide useful means of assessing the reliability of a rating system, which is important in disease diagnosis. Our research was motivated by a study evaluating classification methods for chest radiographs for Pneumoconiosis developed by the International Labor Office in Geneva. The same subjects were evaluated by multiple readers twice using different formats. Focus was on comparing intra-reader reliability of these formats, which, due to the sampling design are correlated. Earlier work in this area dealt with the problem under the scenario that the readers are homogeneous. Our modification offers a Bayesian approach avoiding such simplified assumptions. Simulation studies showed that our model outperforms the frequentist methods in terms of type I error and power even when the rating probabilities differ moderately. We further developed a Bayesian model for comparing dependent agreement measures adjusting for the subject- and rater-level heterogeneity. We adopted a joint analysis that alleviates potential bias stemming from the two-stage method.

162. Estimating an Unknown Multidimensional Prior from Heterogeneous Data via Nonparametric Maximum Likelihood

[IS 16, (page 9)]

Bodhisattva SEN, *Department of Statistics,Columbia University*
Jake SOLOFF, *University of California at Berkeley*
Adityanand GUNTUBOYINA, *University of California at Berkeley*

Statistical inference of stellar populations is complicated by significant observational limitations, in particular, by multivariate, heteroscedastic measurement errors. Empirical Bayes is attractive in such settings, but assumptions about the form of the prior distribution can be hard to justify. We extend the method of nonparametric maximum likelihood (NPMLE) to allow for multivariate and heteroscedastic errors. The NPMLE estimates an arbitrary prior by solving an infinite-dimensional, convex optimization problem; we show that it can be tractably approximated by a finite-dimensional version. The empirical Bayes posterior means have low regret, meaning they closely target the posterior means one would compute with the true prior in hand. Furthermore, the NPMLE can be used for a variety of other purposes such as density estimation and deconvolution. We apply the method to an astronomy data set to construct a fully data driven color-magnitude diagram of 1.4 million stars.

163 . A nonparametric test of cospectrality in networks

[IS 36, (page 17)]

Srijan SENGUPTA, *Statistics,North Carolina State University*

We live in an interconnected world where network valued data arises in many domains, and, fittingly, statistical network analysis has emerged as an active area in the literature. However, the topic of hypothesis testing in networks has received relatively less attention. In this work we consider the problem where one is given two networks, and the goal is to test whether the given networks are cospectral, i.e., they have the same non-zero eigenvalues.

Cospectral graphs have been well studied in graph theory and computer science. Cospectrality is relevant in real-world networks since it implies that the two networks share several important path-based properties, such as the same number of closed walks of any given length, the same epidemic threshold, etc. However, to the extent of our knowledge, there has not been any formal statistical inference work on this topic.

We propose a non-parametric test of co-spectrality by leveraging some recent developments in random matrix theory. We develop two versions of the test one based on an asymptotic bound and one based on bootstrap resampling. We establish theoretical results for the proposed test, and demonstrate its empirical accuracy using synthetic networks sampled from a wide variety of models as well as several well-known real-world network datasets.

This work is in collaboration with Chetkar Jha (University of Pennsylvania) and Indrajit Jana (Indian Institute of Technology, Bhuvaneshwar).

164. Public health data and trend filtering

[IS 47, (page 22)]

James SHARPNACK, *Statistics Department, UC Davis*

We start this talk by outlining the efforts that our group are making to help combat the spread of Covid-19 in our community through the Healthy Davis Together (HDT) project (http://healthydavistogether.org) and in collaboration with the Delphi Lab (http://delphi.cmu.edu). Through a customer discovery process, we identified several data processing and analysis tasks to assist policy makers and scientists including the HDT executive committee, CDPH, and data journalists. Many of the most important tasks identified in this way are not simple prediction tasks, and instead can be cast as spatial clustering, segmentation, change-point localization, and outbreak detection, to name a few. We will highlight one specific problem, that of spatio-temporal segmentation of CA county test positivity proportion using graph trend filtering (GTF). Trend filtering and graph segmentation provide locally adaptive function estimates which can solve a wide range of problems including small area estimation. There are also public policy applications in which the segmentation properties of GTF are desired as they lead to more demonstrably fair and interpretable predictions. We will take a brief tour of GTF, including some of the theoretical advances made in the past several years. A special emphasis will be made on the gaps in our understanding of GTF and what these results tell us about the denoising power of different graph structures. Returning to our application at hand, we apply GTF to Covid-19 test positive proportions of CA counties using a mobility network from the Safegraph mobility data.

165. Statistical Inference for Networks of Point Processes

[IS 11, (page 7)]

Ali SHOJAIE, *Biostatistics, University of Washington*

Advances in calcium florescent imaging have facilitated monitoring of the activity of thousands of neurons in live animals. Data from these live images reveal the firing times of neurons, or their spike trains, in response to various stimuli. I will discuss new methodological, computational and theoretical developments for learning functional connectivity networks from high-dimensional Hawkes processes, which are widely used to model neuronal spike train data. In this talk, we discuss new procedures for learning connectivity networks from high-dimensional point processes and statistical inference procedures for characterizing the uncertainty of the resulting estimators. We also discuss an extension of this procedure to learn networks from multiple experiments, which are commonly used to glean insight into changes in brain connectivity associated with different tasks.

166. Bayesian profiling multiple imputation for missing hemoglobin values in electronic health records

[IS 35, (page 17)]

Yajuan SI, *Survey Research Center, University of Michigan*

Mari PALTA, *University of Wisconsin-Madison*

Maureen SMITH, *University of Wisconsin-Madison*

Electronic health records (EHRs) are increasingly used for clinical and comparative effectiveness research but suffer from missing data. Motivated by health services research on diabetes care, we seek to increase the quality of EHRs by focusing on missing values of longitudinal glycosylated hemoglobin (A1c), a key risk factor for diabetes complications and adverse events. Under the framework of multiple imputation (MI), we propose an individualized Bayesian latent profiling approach to capture A1c measurement trajectories subject to missingness. The proposed method is applied to EHRs of adult patients with diabetes in a large academic Midwestern health system between 2003 and 2013 and had Medicare A and B coverage. We combine MI inferences to evaluate the association of A1c levels with the incidence of acute adverse health events and examine patient heterogeneity across identified patient profiles. We investigate different missingness mechanisms and perform imputation diagnostics. Our approach is computationally efficient and fits flexible models that provide useful clinical insights.

167. Selection of 2-level supersaturated designs for main effects models

[IS 62, (page 28)]

Rakhi SINGH, *Informatics and Analytics,University of North Carolina at Greensboro*

John STUFKEN, *University of North Carolina at Greensboro*

An extensive literature is available on design selection criteria and analysis techniques for 2-level supersaturated designs. The most notable design selection criteria are the popular $E(s^2)$-criterion, $UE(s^2)$-criterion, and Bayes D-optimality criterion, while the most notable analysis technique is the Gauss-Dantzig Selector. It has been observed that while the Gauss-Dantzig Selector is the preferred analysis technique, differences in screening performance of different designs are not captured by any of the common design selection criteria. We develop new design selection criteria inspired by the Gauss-Dantzig Selector, and establish that designs that are better under these criteria also tend to perform better as screening designs.

168. Min–Max Crossover Designs for Two Treatments Binary and Poisson Crossover Trials

[IS 62, (page 28)]

Satya SINGH, *Mathematics,Indian Institute of Technology Hyderabad*

Siuli MUKHOPADHYAY, *Indian Institute of Technology of Bombay*

Harsh RAJ, *Indian Institute of Technology of Hyderabad*

In this talk min–max crossover designs for binary and Poisson crossover trials with two treatments are discussed. Models with and without carryover effects are considered. Min–max designs for periods 2 and 3 are discussed in detail. A sensitivity analysis is performed to assess the robustness of proposed designs when compared to existing optimal designs. An equivalence theorem is provided to verify optimality of min-max designs.

169. An Objective Bayesian Multiple Testing of Binomial Proportions

[IS 58, (page 26)]

Siva SIVAGANESAN, *Division of Statistics and Data Science, Department of Mathemtical Sciences,University of Cincinnati*

Emrah GECILI, *Cincinnati Children's Hospital*

Nilupika HERATH, *University of Cincinnati*

An objective Bayesian approach to multiple testing of (in)equality of two (or more) binomial proportions under different experimental settings is considered. Focusing first on two proportions case, we investigate a selection of priors under the alternative(s) and choose a suitable objective prior that permits certain desirable characteristics. We use simulated and real data sets to compare the results obtained by using this prior with some frequentist approaches. Next, we extend the method to consider multiple testing of ordered proportions, and to allow dependence among the parameters under different settings, and illustrate the approaches using a real data example.

170. Effective use of Visual Analytics in Monitoring Clinical Trial Data

[IS 8, (page 5)]

Abigail SLOAN, *Biostatistics,Pfizer*

Anindita BANERJEE, *Pfizer*

Regular review of ongoing trial data improves data quality and allows early identification of data issues. In this talk, we will describe a variety of plots that can be used to assess blinded data, including waterfall plots, lollipop plots, and longitudinal plots. Such plots can provide intuition when assessing treatment effect and benefit-risk. We will also illustrate tabular displays that can be used to review endpoints composed of multiple components. Indications discussed include inflammatory bowel disease and dermatology indications.

171. Random Effects with Bayesian Additive Regression Trees for Precision Medicine

[IS 23, (page 12)]

Charles SPANBAUER, *Biostatistics,University of Minnesota*

Rodney SPARAPANI, *Medical College of Wisconsin*

Precision medicine is an active area of research that could offer an analytic paradigm shift for clinical trials and the subsequent treatment decisions based on them. Clinical trials are typically analyzed with the intent of discovering beneficial treatments if the same treatment is applied to the entire population under study. But, such a treatment strategy could be suboptimal if subsets of the population exhibit varying treatment effects. Identifying subsets of the population experiencing differential treatment effect and forming individualized treatment rules is a task wellsuited to modern machine learning methods such as treebased ensemble predictive models. Specifically, Bayesian additive regression trees (BART) has shown promise in this regard because of its exceptional performance in outofsample prediction. However, in many cases, the outcome of interest is measured repeatedly on the same subject as in longitudinal data. The traditional BART model assumes independence of the outcomes (conditional on the predictors), so we propose an extension to relax this assumption called mixedBART. We incorporate random effects for longitudinal repeated measures and subject clustering within medical centers. Simulation studies and applications of precision medicine based on real randomized clinical trials data examples are presented.

172. Nonparametric Failure Time with BART and DPM LIO

[IS 23, (page 12)]

Rodney SPARAPANI, *Division of Biostatistics,Medical College of Wisconsin*
Laud, PURUSHOTTAM, *Medical College of Wisconsin*
Logan, BRENT, *Medical College of Wisconsin*
McCulloch, ROBERT, *Arizona State University*
Pratola, MATT, *Ohio State University*

Bayesian Additive Regression Trees (BART) is a nonparametric machine learning method for continuous, dichotomous, categorical and time-to-event outcomes. However, survival analysis with BART currently presents some challenges. Two current approaches each have their pros and cons. Our discrete time approach is free of precarious restrictive assumptions such as proportional hazards and Accelerated Failure Time (AFT), but it becomes increasingly computationally demanding as the sample size increases. Alternatively, a Dirichlet Process Mixture approach is computationally friendly, but it suffers from the AFT assumption. Therefore, we propose to further nonparametrically enhance this latter approach via heteroskedastic BART which will remove the restrictive AFT assumption while maintaining its desirable computational properties.

173. Distributed Bayesian Varying Coefficient Modeling Using a Gaussian Process Prior

[IS 51, (page 23)]

Sanvesh SRIVASTAVA, *Department of Statistics and Actuarial Science,The University of Iowa*
Rajarshi GUHANIYOGI, *UC Santa Cruz*
Cheng LI, *National University of Singapore*
Terrance SAVITSKY, *Bureau of Labor Statistics*

Varying coefficient models (VCMs) are widely used for estimating nonlinear regression functions for functional data. Their Bayesian variants using Gaussian process priors on the functional coefficients, however, have received limited attention in massive data applications, mainly due to the prohibitively slow posterior computations using Markov chain Monte Carlo (MCMC) algorithms. We address this problem using a divide-and-conquer Bayesian approach. We first create a large number of data subsamples with much smaller sizes. Then, we formulate the VCM as a linear mixed-effects model and develop a data augmentation algorithm for obtaining MCMC draws on all the subsets in parallel. Finally, we aggregate the MCMC-based estimates of subset posteriors into a single Aggregated Monte Carlo (AMC) posterior, which is used as a computationally efficient alternative to the true posterior distribution. Theoretically, we derive minimax optimal posterior convergence rates for the AMC posteriors of both the varying coefficients and the mean regression function. We provide quantification on the orders of subset sample sizes and the number of subsets. The empirical results show that the combination schemes that satisfy our theoretical assumptions, including the AMC posterior, have better estimation performance than their main competitors across diverse simulations and in a real data analysis.

174. Musings about supersaturated designs

[IS 50, (page 23)]

John STUFKEN, *Informatics and Analytics,University of North Carolina at Greensboro*
Rakhi SINGH, *University of North Carolina at Greensboro*

Screening designs are used in design of experiments when, with limited resources, important fac-

tors are to be identified from a large pool of factors. Typically, a screening experiment will be followed by a second experiment to study the effect of the identified factors in more detail. As a result, the screening experiment should ideally screen out a large number of factors to make the follow-up experiment manageable, without screening out important factors.

The Gauss-Dantzig Selector (GDS) is often the preferred analysis method for screening designs. While there is ample empirical evidence that fitting a main-effects model can lead to incorrect conclusions about the factors if there are interactions, including two-factor interactions in the model increases the number of model terms dramatically and challenges the GDS analysis. We discuss a new analysis method, called Gauss Dantzig Selector Aggregation over Random Models (GDS-ARM), which aggregates the effects from different iterations of the GDS analysis performed using different sets of randomly selected interactions columns each time.

175. Bayesian Analysis of the Covariance Matrix of a Large Dimensional Multivariate Normal Distribution with Shrinkage Inverse Wishart Priors

[IS 45, (page 21)]

Dongchu SUN, *Statistics,University of Nebraska-Lincoln*

James O. BERGER, *Duke University*

Chengyuan SONG, *East China NOrmal University*

Modern statistical methods allow us to analyzing more and more complicated data.

Real data sets grow both in size and complexity. For example, The Cancer Genome Atlas (TCGA) contains a large number of heterogenous data sets obtained by measuring different genomic phenomena, such as gene expression, copy number, snips, etc., on the same tissue samples.

Meantime, we often face the challenge to discover the structure for a large number unknown parameters with a few observations.

Bayesian analysis for the covariance matrix of a multivariate normals has always received a lot of attention, especially for a large dimensional case. We propose a new class of priors for the covariance matrix, including both inverse Wishart and reference priors as special cases. The main motivation for the new class is to have available priors –both subjective and objective– that do not “force eigenvalues apart,” which is a criticism of inverse Wishart and the Jeffreys priors. Extensive comparison of these ‘shrinkage priors’ with inverse Wishart and Jeffreys priors is undertaken, with the new priors seeming to have considerably better performance. A number of curious facts about the new priors are also observed, such as that the low rank learning, in the sense that posterior distribution will be proper with just three vector observations from the multivariate normal distribution –regardless of the dimension of the covariance matrix – and that useful inference about features of the covariance matrix can be possible. Finally, a new MCMC algorithm is developed for this class of priors and is shown to be computationally effective for matrices over 100 dimensions.

176. Missing data: sensitivity analysis or supplementary analysis?

[IS 49, (page 22)]

Steven SUN, *Statistical Decision Science,Janssen R&D*

Missing data represent a potential source of bias in a clinical trial. In some trials missing data for some subjects is inevitable by design, such as administrative censoring in trials with survival outcomes. Sensitivity analyses are often used to handle missing data. The addendum to the ICH E9 guideline finalized in November 2019 introduces the estimand framework and emphasizes the difference between sensitivity analyses and supplementary analyses. In this talk, we will discuss some common statistical methods for handing missing data under the estimand framework. In particular, we will illustrate how we should classify them as sensitivity analyses or supplementary analyses.

177 . Improved Nonparametric Empirical Bayes Estimation By Transfer Learning

[IS 16, (page 9)]

Wenguang SUN, *University of Southern California,Session on Empirical Bayes Methodology*

Gourab MUKHERJEE, *University of Southern California*

Jiajun LUO, *University of Southern California*

In this talk I discuss a class of nonparametric integrative Tweedie (NIT) estimators for empirical Bayes data-sharing shrinkage estimation. When applied in conjunction with reproducing kernels and convex optimization techniques, NIT provides superior and robust performance and scales well with growing number of parameters. The new estimation framework is capable of handing multiple and possibly correlated auxiliary sequences and is flexible for incorporating various structural constraints into the data-driven decision rule. We develop theory to establish the convergence rates for the risk of the data-driven NIT. The theory provides important insights on the benefits and caveats of utilizing multivariate auxiliary data. Numerical studies show that our approach achieves substantial gain in empirical performance over existing methods in many settings. Joint work with Jiajun Luo and Gourab Mukherjee.

178. Bias-Variance Tradeoffs in Joint Spectral Embeddings

[IS 54, (page 24)]

Daniel SUSSMAN, *Mathematics and Statistics,Boston University*
Benjamin DRAVES, *Boston University*

We consider the ramifications of utilizing biased latent position estimates in subsequent statistical analysis in exchange for sizable variance reductions in finite networks. We establish an explicit bias-variance tradeoff for latent position estimates produced by the omnibus embedding in the presence of heterogeneous network data. We reveal an analytic bias expression, derive a uniform concentration bound on the residual term, and prove a central limit theorem characterizing the distributional properties of these estimates.

179. Variable selection in mixture of regression models: Uncovering cluster structure and relevant features

[IS 51, (page 23)]

Mahlet TADESSE, *Department of Mathematics and Statistics,Georgetown University*

Identifying latent classes and component-specific relevant predictors can shed important insights when analyzing data. In this talk, I will present methods we have proposed to address this problem in a unified manner bycombining ideas of mixture of regression models and variable selection.. These include (1) a stochastic partitioning method to relate two high-dimensional datasets, (2) a penalized mixture of multivariate generalized linear regression models, and (3) a mixture of regression trees approach.I will illustrate the methods with various applications.

180. TITE-BOIN-ET: Time-to-event Bayesian optimal interval design to accelerate dose-finding based on both efficacy and toxicity outcomes

[IS 26, (page 13)]

Kentaro TAKEDA, *Data Science,Astellas Pharma Global Development, Inc*
Satoshi MORITA, *Kyoto University*
Masataka TAGURI, *Yokohama City University*

The primary goal of a dose-finding trial for the novel anticancer agents is to identify an optimal dose (OD), defined as the tolerable dose having adequate efficacy under the unpredictable dose-toxicity and dose-efficacy relationships. It is also quite important to accelerate early stage trials to shorten the entire period of drug development. 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 nonparametric and a model-assisted design. A simulation study shows that the TITE-BOIN-ET design has advantages compared 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.

181 . Group Sequential Holm and Hochberg Procedures

[IS 43, (page 20)]

Ajit TAMHANE, *Ajit Tamhane,Northwestern University*
Jiangtao GOU, *Villanova University*
Dong XI, *Novartis Pharmaceuticals*

The problem of testing multiple hypotheses using a group sequential procedure (GSP) arises often in clinical trials. We review several group sequential Holm (GSHM) procedures proposed in the literature and clarify the relationships between them. In particular, we show which procedures are equivalent or, if different, which are more powerful and what are their pros and cons. In analogy with the relationship between the fixed sample Hochberg procedure as a reverse (step-up) application of the fixed sample Holm (step-down) procedure, we propose a group sequential Hochberg (GSHC) procedure as a reverse application of a particular GSHM procedure. We conducted an extensive simulation study to evaluate the familywise error rate (FWER) control of two GSHM and the GSHC procedures and to compare their powers. All procedures are illustrated with a common numerical example for which the data are chosen to bring out the differences between them. A real case study is presented to illustrate application of these procedures.

182. Bayesian Emax dose response modeling

[IS 32, (page 16)]

Neal THOMAS, *Pfizer, Groton CT USA,Pfizer*

The design and analysis of dose response based on Bayesian Emax models will be described. Meta-analyses of a large number of dose response studies will be briefly described to provide an empirical basis for selection of a specific parametric model that accurately describes a high proportion of dose response curves. Analyses of these studies also provide an empirical basis for a prior distribution for the model parameters when combined with information that is available at the time most dose response studies are designed. Software (R package clinDR) will be described that supports implementation of Emax models for clinical data. An example illustrating the methods and software will be presented.

183. Optimal Design Theory in Early-Phase Dose-Finding Problems with Late-onset Toxicity

[IS 33, (page 16)]

Tian TIAN, *Biostatistics,BeiGene*
Min YANG, *University of Illinois at Chicago*

In oncology clinical trials, early-phase studies usually are conducted as dose-finding studies and the dose to be found here is often the one with the greatest therapeutic effect and acceptable toxicity. We often refer to this dose as the MTD, i.e., maximum tolerated dose. A great deal of methods have been proposed to address the MTD estimation problem, among which the CRM (continual reassessment method) stands out due to its simplicity and outstanding performance. We extend the classic CRM by incorporating the idea of optimal design (OD) theory. We denote the new method as OD-CRM. In the talk we will present some theoretical results as well as computational properties of the new method.

184. Correcting population stratification in association studies of rare genetic variants using generalized PCA

[IS 37, (page 17)]

Asuman TURKMEN, *Department of Statistics,OHIO STATE UNIVERSITY*
Nedret BILLOR, *Auburn University*
Yuan YUAN, *Auburn University*

Population stratification has been the focus of many recent studies that evidenced allele frequency differences between populations. Unaccounted population stratification can lead to false-positive findings and can mask the true association signals in identification of disease-related genetic variants. Recent studies showed that rare variants can show a stratification that is systematically different from, and typically stronger than, common variants, and this is not necessarily corrected by existing methods. In this study, we investigate performances of different variants of principal component analysis (PCA) methods including generalized PCA and similarity-matrix-based PCA to detect underlying structures for rare and common variants. Our results indicated that generalized PCA (i.e., logistic PCA) can adjust for population stratification in rare variants much more effectively than standard PCA while their performances are comparable for common variants.

185 . An empirical Bayes approach to estimating dynamic models of co-regulated gene expression

[Student Paper Competition 2, (page 20)]

Sara VENKATRAMAN, *Department of Statistics and Data Science,Cornell University, Department of Statistics and Data Science*
Sumanta BASU, *Cornell University, Department of Statistics and Data Science*
Myung Hee LEE, *Weill Cornell Medical College, Department of Medicine*
Martin T. WELLS, *Cornell University, Department of Statistics and Data Science*

Time-course gene expression datasets provide insight into the dynamics of complex biological processes, such as immune response and organ development. It is of interest to identify genes with similar temporal expression patterns because such genes are often biologically related. However, this task is complicated by the high dimensionality of genetic datasets and the nonlinearity of gene expression time dynamics. We propose an empirical Bayes approach to estimating ordinary differential equation (ODE) models of gene expression, from which we derive metrics that capture similarities in the time dynamics of two genes. These metrics, which we call the Bayesian lead-lag R^2 values, can be used to construct clusters or networks of functionally-related genes. The key feature of our method is that it leverages biological databases that document known interactions between genes. This information is automatically used to define informative prior distributions on the ODE models parameters. We then derive data-driven shrinkage parameters from Steins unbiased risk estimate that optimally balance the ODE models fit to both the data and external biological information. Using real gene expression data, we demonstrate that our biologically-informed similarity metrics allow us to recover sparse, interpretable gene networks. These networks reveal new insights about the dynamics of biological systems.

186. New Perspectives on Causal Inference in Clinical Trials With Multiple Endpoints, Repeated Measures and Subject Non-Adherence

[IS 1, (page 3)]

Hakeem WAHAB, *Statistics,Purdue University*
Stephen RUBERG, *Analytix Thinking*
Hege MICHIELS, *Ghent University*
Arman SABBAGHI, *Purdue University*

Randomized clinical trials will inevitably have some patients whose outcomes are missing due to complications such as adverse reactions, lack of efficacy and excess efficacy from treatment. These disruptions or intercurrent events to the planned trial protocol can muddle or confound the effect of both experimental and control treatments under investigation. Specifically, such complications yield latent strata or subpopulations of patients characterized by their treatment adherence or compliance behaviors. These strata must be taken into account in order to obtain valid causal inferences on the effect of the receipt of treatment, and not merely the assignment of treatment. To that end, we developed a Data Generating Model embedded in an Rshiny app that simulates clinical trials under the Rubin Causal Model that captures the treatment causal effect while accounting for intercurrent events in clinical trials with multiple endpoints. This app enables users to control patient compliance through different sources of discontinuity with varying functional trends, and understand operating characteristics of treatment effect estimators obtained by different models for various estimands.

187. Sampling for Massive Data with Rare Events

[IS 57, (page 26)]

Hai Ying WANG, *Statistics,University of Connecticut*

This paper studies binary regression for rare events data, or imbalanced data, where the number of events (observations in one class, often called cases) is significantly smaller than the number of nonevents (observations in the other class, often called controls). We first derive the asymptotic distribution of the maximum likelihood estimator (MLE) of the unknown parameter, which shows that the asymptotic variance convergences to zero in a rate of the inverse of the number of the events instead of the inverse of the full data sample size. This indicates that the available information in rare events data is at the scale of the number of events instead of the full data sample size. Furthermore, we prove that sub-sampling a small proportion of the nonevents, the resulting sub-sampled estimator may have identical asymptotic distribution to the full data MLE. This demonstrates the advantage of sub-sampling nonevents for rare events data, because this procedure significantly reduces the computation and/or data collection costs. Another common practice in analyzing rare events data is to over-sample (replicate) the events, which has a higher computational cost. We show that this procedure may even result in efficiency loss in terms of parameter estimation.

188 . Analytic framework for non-randomized single-arm clinical trials with external RWD control

[IS 7, (page 5)]

Hongwei WANG, *Global Medical Affairs Statistics,AbbVie*
Yixin FANG, *AbbVie*
Weili HE, *AbbVie*

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.

189 . Causal inference in Mendelian Randomization with weak and heterogenous instruments

[IS 28, (page 14)]

Jingshu WANG, *Statistics,University of Chicago*
Qingyuan ZHAO, *University of Cambridge*
Jack BOWDEN, *The University of Exeter*
Dylan SMALL, *University of Pennsylvania*
Nancy R. ZHANG, *University of Pennsylvania*

Mendelian randomization (MR) is a method using genetic variants as instruments to estimate the causal effects of risk factors in presence of unmeasured confounding. We derive a maximum profile likelihood estimator with provable consistency and asymptotic normality to accommodate the inclusion of large numbers of weakly associated instruments. To deal with both systematic and idiosyncratic pleiotropy in MR, we propose a consistent and asymptotically normal estimator by robustifying and adjusting the profile score. In addition, we also develop a comprehensive framework that can diagnose the existence of multiple pleiotropic pathways that violates our model assumptions, identify causal directions and adjust for confounding risk factors. With the new framework, 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.

190. Generalized orthogonal subsampling for predictive stability

[IS 57, (page 26)]

Lin WANG, *Department of Statistics,George Washington University*
Yi ZHANG, *George Washington University*

Training and testing data commonly have different distributions due to sampling bias or domain shift. Predictive stability refers to the merit of a trained model that makes accurate predictions on unknown testing data whose distributions may vary from that of the training data. We propose a generalized orthogonal subsampling approach to efficiently establish stable models from big data. The approach is inspired by the fact that an orthogonal array provides the universally optimal experimental design and the minimum worst prediction error for fixed-effect models. Theoretical results show that the proposed approach reduces the confounding between covariates, enables accurate estimations of causal effects, and therefore enhances model stability. The trained model on the selected subsample is also robust to model misspecification. The stability of the trained model is demonstrated through numerical analysis of both synthetic and real data.

191. Practical considerations of using historical data in clinical trials: when, what and how much do we borrow?

[IS 32, (page 15)]

Ling WANG, *Worldwide Research, Development and Medical,Pfizer*

In early clinical development, historical data are frequently used in study designs to borrow information, reduce sample size and minimize patient's exposure to placebo. There is a wealth of Bayesian methods available for such trial designs. In this talk, I will discuss the practical aspects of should we use historical data or not, designing and implementing trials using historical data and different methods of borrowing with pros and cons.

192. Practical considerations of using historical data in clinical trials: when, what and how much do we borrow?

[Student Paper Competition 2, (page 20)]

Selena WANG, *Ohio State University*
Subhadeep PAUL, *Ohio State University*

In many application problems in social, behavioral, and economic sciences, researchers often have data on a social network among a group of individuals along with attributes or behaviors for each individual. The attributes are often measured as responses to survey questionnaire or relational data based on observed choices. The attributes are therefore often high-dimensional. To analyze such data structures, we propose a joint Latent Space Model (JLSM) that summarizes information from the social network and the attribute information in a person-attribute joint latent space. We develop a Variational Bayesian Expectation-Maximization estimation algorithm to estimate the attribute and person locations in the joint latent space. This methodology allows for effective integration, informative visualization, and prediction of social networks and attribute information. Using JLSM, we analyze user networks and behaviors in multi-modal social media systems like Instagram and YouTube. We also explore the French financial elites based on their social networks and their career, political views, and social status. We observe a division in the social circles of the French elites in accordance with the differences in their individual characteristics. A R package jlsm is developed to fit the models proposed in this paper and is publicly available from the CRAN repository https://cran.r-project.org/web/packages/jlsm/jlsm.pdf.

193. Statistical Considerations in Clinical Trial Design for Rare Diseases

[IS 19, (page 10)]

Zailong WANG, *AbbVie,AbbVie*
Lanju ZHANG, *AbbVie*

The FDA published DRAFT GUIDANCE for Industry regarding Rare Diseases: Common Issues in Drug Development in January 2019. FDA regulations provide flexibility in applying regulatory standards because of the many types and intended uses of drugs for rare diseases. In this presentation, the main points from FDA guidance will be summarized, including natural history studies, non-clinical studies, endpoints and sample size consideration, and clinical trial design strategies. For illustration purpose, an example is presented in application to FDA guidance for primary Sjogrens Syndrome disease (pSS). Detailed strategy for modifying assessment measures as primary efficacy endpoints, power, and sample size consideration, new pSS study design and analysis strategies as historical data incorporation and futility analysis are also presented.

Key words: rare disease, efficacy endpoints, power and sample size, historical data, futility analysis

Support of this publication was provided by AbbVie. AbbVie participated in the review and approval of the content. Zailong Wang and Lanju Zhang are employees of AbbVie Inc. and may own AbbVie stock.

194. Mixed-effects location scale modeling for the analysis of accelerometry data

[IS 60, (page 27)]

Whitney WELCH, *Department of Preventive Medicine,Northwestern University Feinberg School of Medicine*
Donald HEDEKER, *University of Chicago*
Bonnie SPRING, *Northwestern University Feinberg School of Medicine*
Juned SIDDIQUE, *Northwestern University Feinberg School of Medicine*

Purpose: To introduce a statistical technique, the mixed-effects location scale model, for analysis of longitudinal accelerometer-based physical activity (PA) data. This approach jointly models both the mean (location) and within-subject variability (scale) of participants PA over time as a function of covariates, since within-person variability may be an important construct to explore in PA interventions. Random effects are included in both models to allow for subject-specific deviations beyond the effect of covariates. These random effects can be correlated. Methods: Participants (N=204, 77% female, age=33±11y, BMI=28.2±7.1 kg/m2) in the Make Better Choices Study were randomized to one of two activity-related intervention arms: 1) increase moderate-to-vigorous PA (MVPA) (PA group) or 2) decrease sedentary active control (SB group). Physical activity was measured by accelerometer for 5 weeks: a 2 week baseline assessment phase and a 3 week intervention follow-up phase: week 1 (rx1) and weeks 2 and 3 (rx23). The outcome MVPA min/d was analyzed using the mixed-effects location scale model in the MIXREGLS software program in STATA. Results: The mean model shows a significant group by time interaction (MVPA group by rx1: B=6.32 (95%CI: 3.93, 8.7) MVPA group by rx23: B=9.85 (95% CI: 7.59, 12.10)) indicating that those in the PA group had significantly greater MVPA min/d at rx1 and rx23 compared to the SB group. The PA group by rx23 interaction was significant in the within-subject variance model, suggesting that those in the PA group had significantly more variability in MVPA min/d during follow-up phase rx23 compared to the SB group. The random-location effect is positively associated with the within subject variance, participants with higher mean min/d MVPA tend to have higher min/d MVPA variability (=0.70 (95% CI: 0.60, 0.80)). The scale standard deviation is significant indicating that some participants MVPA min/d are significantly more dispersed than other participants even after adjusting for group and time effects (=0.60 (95% CI: 0.55, 0.64)). Conclusions: The location-scale mixed model provides a new approach for examining the mean and variability of min/d of MVPA in longitudinal data. To demonstrate, we applied this model to a randomized controlled trial to increase PA in inactive adults.

195. Novel strategy for disease risk prediction incorporating predicted gene expression and DNA methylation: a multi-phased study of prostate cancer

[IS 65, (page 29)]

Chong WU, *Department of Statistics, Florida State University,Department of Statistics, Florida State University*
Jingjing ZHU, *Cancer Epidemiology Division, Population Sciences in the Pacific Program, University of Hawaii Cancer Center, University of Hawaii at Manoa, Honolulu, HI, USA*
Austin KING, *Department of Statistics, Florida State University*
Xiaoran TONG, *Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA*
Qing Lu, Jong Y Park, Liang Wang, Guimin Gao, Hong-Wen Deng, Yaohua Yang, Karen E Knudsen, Timothy R Rebbeck, Jirong Long, Wei Zheng, Wei Pan, David V Conti, Christopher A Haiman, Lang WU,

Prostate cancer (PCa) is known to be highly heritable and polygenic, indicating that polygenic risk scores (PRS) have great promise in identifying males at risk. Most existing PRS are constructed by summing the cumulative effect of common genetic variants. DNA methylation and gene expression are known to play important roles in PCa etiology, however, it has not yet been possible to incorporate information of DNA methylation and gene expression into PRS. Here, we develop and validate an improved PRS for PCa risk by incorporating genetically predicted gene expression and DNA methylation and other genomic information using an integrative method. Our constructed PRS has improved performance (C statistics: 76.1%) over PRS constructed by individual benchmark methods. Furthermore, our new PRS has much higher risk assessment power than family history. The overall net reclassification improvement was 42.7% for adding PRS to the baseline model compared with 12.5% for adding family history. Our novel method can be applied to other complex traits to improve their risk prediction.

196. Advanced Machine Learning to Predict Enrollment to Expedite Clinical Trials

[IS 7, (page 5)]

Yunzhao XING, *Data and Statistical Science,AbbVie*
Li WANG, *AbbVie*

In clinical development, other than probability of success, speed is crucial. However, from the article in Clinicaltrialsarena.com, almost 80% of clinical trials fail to meet enrollment timeline. Agile and efficient patient recruitment strategy becomes one of the critical factors for successful clinical trials planning and execution. Multiple factors, including study countries, clinical sites, primary investigators, et. al, impact the patient recruitment rate. The correlation between those factors and patient recruitment, however, is usually ambiguous and complicated. Machine learning is an excellent technique to extract information from complex factors and it provides a non-linear predictive model. This presentation will illustrate a general framework applying machine learning to predict and potentially accelerate patient recruitment in clinical trials.

The following will be included in the presentation This presentation was sponsored by AbbVie. AbbVie contributed to the design, research, and interpretation of data, writing, reviewing, and approving the publication. All authors are employees of AbbVie Inc. and may own AbbVie stock.

197. Semiparametric Bayesian Markov Analysis of Personalized Benefit-Risk Assessment

[IS 52, (page 23)]

Dongyan YAN, *Discovery & Development Statistics,Eli Lilly and Company*
Subharup GUHA, *Department of Biostatistics, University of Florida*
Chul AHN, *Division of Biostatistics, Center for Devices and Radiological Health, Office Surveillance and Biometrics, U.S. Food and Drug Administration*

Ram TIWARI, *Statistical Methodology at Bristol Myers Squibb*

The development of systematic and structured approaches to assess benefit-risk of medical products is a major challenge for regulatory decision makers. Existing benefit-risk methods depend only on the frequencies of mutually exclusive and exhaustive categories in which the subjects fall, and the responses of individuals are allowed to belong to any of the other categories during their post-withdrawal visits. In this article, we introduce a semiparametric Bayesian Markov model (SBMM) that treats the withdrawal category as an absorbing state, and analyzes subject-level data for multiple visits, accounting for any within-patient dependencies in the response profiles. A log-odds ratio model is used to model the subject-level effects by assuming a ratio of transition probabilities, with respect to a "reference" category. A Dirichlet process is used as a semiparametric model for the subject-level effects to flexibly capture the underlying distributions of the personalized response profiles without making strong parametric assumptions. This also allows the borrowing of strength between the patients and achieves dimension reduction by allocating similar response profiles patterns into an unknown number of latent clusters. We analyze a motivating clinical trial dataset to assess the personalized benefit-risks in each arm and evaluate the aggregated benefits and risks associated with the drug Exalgo.

198 . Score-Matching Representative Approach for Big Data Analysis with Generalized Linear Models

[IS 3, (page 4)]

Jie YANG, *Department of Mathematics, Statistics, and Computer Science,University of Illinois at Chicago*
Keren LI, *Northwestern University*

We propose a fast and efficient strategy, called the representative approach, for big data analysis with generalized linear models, especially for distributed data with localization requirements or limited network bandwidth. With a given partition of massive dataset, this approach constructs a representative data point for each data block and fits the target model using the representative dataset. In terms of time complexity, it is as fast as the subsampling approaches in the literature. As for efficiency, its accuracy in estimating parameters given a homogeneous partition is comparable with the divide-and-conquer method. Supported by comprehensive simulation studies and theoretical justifications, we conclude that mean representatives (MR) work fine for linear models or generalized linear models with a flat inverse link function and moderate coefficients of continuous predictors. For general cases, we recommend the proposed score-matching representatives (SMR), which may improve the accuracy of estimators significantly by matching the score function values. As an illustrative application to the Airline on-time performance data, we show that the MR and SMR estimates are as good as the full data estimate when available.

199. Proof of concept and dose estimation with binary responses

[IS 33, (page 16)]

Min YANG, *Department of Mathematics, Statistics, and Computer Science,University of Illinois at Chicago*

POC and dose estimation are two critical steps in drug development. Hybrid approaches of combining multiple testing and modeling have been proved to be more effective than pairwise comparisons. While there are considerable efforts on the improving the efficiency of POC test and subsequent target dose estimation, there lack of systematic investigation on implement of hybrid approaches when the response is binary. Focusing on binary response, we study the hybrid approaches systematically in the following three aspects: (i) How to select the candidate set of plausible dose response models? (ii) How to allocate the sample size across the dose levels? (iii) Is MCP-MOD suitable for binary responses? (iv) Should we choose logit link function for binary response? and (v) What estimation method should be used for target dose estimation?

200. Fast and accurate computation of large-scale kernel ridge regression

[IS 40, (page 18)]

Yun YANG, *Statistics,University of Illinois Urbana-Champaign*

Nystrom approximation is a fast randomized method that rapidly solves kernel ridge regression (KRR) problems through sub-sampling the n-by-n empirical kernel matrix appearing in the objective function. However, the performance of such a subsampling method heavily relies on correctly estimating the statistical leverage scores for forming the sampling distribution, which can be as costly as solving the original KRR. In this work, we propose a linear time (modulo poly-log terms) algorithm to accurately approximate the statistical leverage scores in the stationary-kernel-based KRR with theoretical guarantees. Particularly, by analyzing the first-order condition of the KRR objective, we derive an analytic formula, which depends on both the input distribution and the spectral density of stationary kernels, for capturing the non-uniformity of the statistical leverage scores. Numerical experiments demonstrate that with the same prediction accuracy our method is orders of magnitude more efficient than existing methods in selecting the representative sub-samples in the Nystrom approximation.

201 . Statistical Considerations in Biomarker Cutoff Determination, Development and Validation in Immuno-Oncology Studies

[IS 55, (page 25)]

Jiabu YE, *Oncology Biometrics AstraZeneca,AstraZeneca*
Feng LIU, *Oncology Biometrics AstraZeneca*

While treatment with immune checkpoint inhibitors with or without chemotherapy has transformed the landscape of non-small cell lung cancer management, a substantial proportion of patients do not experience sustained survival benefit. Appropriate biomarkers can help identify patients likely to derive benefit from various therapeutic regimens. This is more challenging where there is heterogeneity in patient characteristics, and where clinical resistance has been established eg in relapsed metastatic disease. Tumor mutational burden (TMB) has been shown to be associated with survival benefit in patients with nonsmall cell lung cancer (NSCLC) treated with immune checkpoint inhibitors. Measuring TMB in the blood (bTMB) using circulating cell-free tumor DNA (ctDNA) would offer practical advantages compared with TMB measurement in tissue (tTMB); however, there is a need for validated assays and identification of optimal cutoffs. We applied minimal p value cross-validation approaches based on the phase 3 MYSTIC trial of durvalumab (anti-PD-L1) ± tremelimumab (anti-CTLA-4) for first-line treatment of metastatic NSCLC. These cutoffs were assessed by their HR distributions among test sets. The minimal P value cross-validation approach confirmed the selection of bTMB $\geq$ 20 (mut/Mb) as the optimal cutoff for clinical benefit with durvalumab + tremelimumab in the MYSTIC study and is shown to be a robust statistical method for determining the optimal bTMB cutoff.

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

[IS 33, (page 16)]

Ying YUAN, *Biostatistics,University of Texas MD Anderson Cancer Center*
Ruitao LIN, *University of Texas MD Anderson Cancer Center*
Yanhong ZHOU, *University of Texas MD Anderson Cancer Center*
Fangrong YAN, *China Pharmaceutical University*
Daniel LI, *Bristol-Myers Squibb*

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 pretabulated 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 to facilitate the application of the BOIN12 design.

203. Bayesian Basket Trial Design with False Discovery Rate Control

[IS 21, (page 11)]

Emily ZABOR, *Department of Quantitative Health Sciences, Cleveland Clinic*
Brian HOBBS, *The University of Texas at Austin*
Michael KANE, *Yale University*
Satrajit ROYCHOUDHURY, *Pfizer, Inc*
Lei NI, *U.S. Food and Drug Administration*

Oncology therapies have evolved over time, from cocktails of cytotoxic drugs to non-cytotoxic therapies that target specific pathways in tumor cells or promote anti-cancer immunity. Drug developers endeavor to understand treatment benefit heterogeneity at the patient level and establish actionable targets through biomarker-guided therapies. Recent advances in developing tumor agnostic treatments have identified molecular targets that define patient subpopulations in a manner that supersedes conventional criteria for cancer classification. These successes have produced effective targeted therapies that are administered to patients regardless of their tumor histology. Trials have evolved as well with master protocol designs. By blending translational and clinical science, basket trials in particular are well-suited to investigate and develop targeted therapies among multiple cancer histologies. This paper provides guidance for design and analysis calibration of basket trials based on the multisource exchangeability model. It also introduces an approach devised to design basket trials for control of the false-discovery rate. Additionally, the methodology is applied to dissect the heterogeneity of the SUMMIT trial.

204. A practical Response Adaptive Block Randomization (RABR) design with analytic type I error protection

[IS 2, (page 3)]

Tianyu ZHAN, *Data and Statistical Sciences, AbbVie Inc.,Data and Statistical Sciences, AbbVie Inc.*

Response adaptive randomization (RAR) is appealing from methodological, ethical, and pragmatic perspectives in the sense that subjects are more likely to be randomized to better performing treatment groups based on accumulating data. However, applications of RAR in confirmatory drug clinical trials are limited largely due to its lack of control of randomization ratios to different treatment groups per regulatory requirements and protocol pre-specifications. To address this issue, we propose a Response Adaptive Block Randomization (RABR) design allowing arbitrarily pre-specified randomization ratios for the control and high-performing groups to meet clinical trial objectives. While taking advantage of RAR to assign more subjects to more effective treatments, our proposed method targets at achieving desired numbers of subjects in the final selected treatment and control groups, meeting regulatory requirements. Using the sample size adaptive design technique requiring no complex stochastic process arguments, we demonstrate that our method allows the conventional analysis with a controlled type I error rate. The advantages of the proposed RABR in terms of robustly controlling randomization ratios and increasing statistical power are clearly shown as compared with the popular Doubly Adaptive Biased Coin Design (DBCD) via statistical simulations and a practical clinical trial design example.

205. An Optimal Statistical and Computational Framework for Generalized Tensor Estimation

[IS 63, (page 28)]

Anru ZHANG, *Statistics,University of Wisconsin-Madison / Duke University*

The analysis of tensor data has become an active research topic in statistics and data science recently. This paper describes a flexible framework for generalized low-rank tensor estimation problems that includes many important instances arising from applications in computational imaging, genomics, and network analysis. The proposed estimator consists of finding a low-rank tensor fit to the data under generalized parametric models. To overcome the difficulty of non-convexity in these problems, we introduce a unified approach of projected gradient descent that adapts to the underlying low-rank structure. Under mild conditions on the loss function, we establish both an upper bound on statistical error and the linear rate of computational convergence through a general deterministic analysis. Then we further consider a suite of generalized tensor estimation problems, including sub-Gaussian tensor denoising, tensor regression, and Poisson and binomial tensor PCA. We prove that the proposed algorithm achieves the minimax optimal rate of convergence in estimation error. Finally, we demonstrate the superiority of the proposed framework via extensive experiments on both simulated and real data.

206. On the Implementation of Robust Meta-Analytical-Predictive Prior

[IS 46, (page 21)]

Hongtao ZHANG, *Global Biometrics and Data Sciences,Bristol Myers Squibb*
Alan Y CHIANG, *Bristol Myers Squibb*
Mike BRANSON, *UCB Pharma*

Prospectively leveraging historical control information has become increasingly popular in biometrics research. For placebo or drugs that are already being used widely, such historical controls may provide useful prior information. However, robustness of analysis in the presence of prior-data conflicts is arguably an important topic. Meta-analytical-predictive (MAP) prior is a well-known dynamic borrowing method for this purpose. A robust MAP (rMAP) prior has been shown to improve the robustness by adding a weakly informative component to the original MAP prior. Specific implementations of rMAP can vary depending on where the weakly informative prior is integrated to the relevant historical data. In this manuscript, we outline the differences among three approaches and design simulation studies in which their performances are compared. Our results suggest that their performances can differ considerably, contrary to the perception that they are equivalent in the robustification setting.

207. Bi-level graphical modeling of functional connectivity analysis of resting-state fMRI data

[IS 25, (page 12)]

Lin ZHANG, *Division of Biostatistics, University of Minnesota*

We consider a novel problem, bi-level graphical modeling, in which multiple individual graphical models can be considered as variants of a common group-level graphical model and inference of both the group- and individual-level graphical models are of interest. We propose a novel random covariance model to learn the group- and individual-level graphical models simultaneously with a new measure of degrees-of-freedom for model complexity that is useful for model selection. We apply the method to our motivating clinical data, a multi-subject resting-state fMRI dataset collected from participants diagnosed with schizophrenia, identifying both individual- and group-level graphical models of functional connectivity. We further extend the method for covariance-based clustering to account for heterogeneity commonly present among multi-subject fMRI data.

208. Element-wise estimation error of a total variation regularized estimator for change point detection

[IS 47, (page 22)]

Teng ZHANG, *Teng Zhang,University of Central Florida*

This work studies the total variation regularized l2 estimator (fused lasso) in the setting of a change point detection problem. Compared with existing works that focus on the sum of squared estimation errors, we give bound on the element-wise estimation error. Our bound is nearly optimal in the sense that the sum of squared error matches the best existing result, up to a logarithmic factor. This analysis of the element-wise estimation error allows a screening method that can approximately detect all the change points. We also generalize this method to the muitivariate setting, i.e., to the problem of group fused lasso.

209. Edgeworth expansion for network moments

[IS 63, (page 28)]

Yuan ZHANG, *Statistics,The Ohio State University*
Dong XIA, *Hong Kong University of Science and Technology*

The aim is to derive and use high-order expansions of network moment statistics for exchangeable network models, including the popular stochastic block model for one- and two-sample network inferences, under very mild assumptions. By this approach, we can achieve the following two goals simultaneously (i) higher-order control of the type I error; and (ii) rate-optimal separation condition on the alternative hypothesis for the test to be consistent. Notice that goal (i) was previously only achieved by computationally expensive bootstrap methods with no power guarantees. We also demonstrate our approach's effectiveness in numerical examples.

210. A regression modeling approach to simultaneous estimation

[IS 53, (page 24)]

Dave ZHAO, *Statistics,University of Illinois at Urbana-Champaign*

Simultaneous estimation problems have a long history in statistics and have become especially common and important in genomics research: modern technologies can simultaneously assay tens of thousands to even millions of genomic features that can each introduce an unknown parameter of interest. These applications reveal some conceptual and methodological gaps in the standard empirical Bayes approach to simultaneous estimation. This talk proposes that some of these challenges can be resolved by adopting an alternative approach based on regression modeling, and will illustrate several estimators that perform well in theory and practice in complex settings.

211. A Log–Linear Model for Inference on Bias in Microbiome Studies

[IS 12, (page 7)]

Ni ZHAO, *Johns Hopkins University,assistant professor of Biostatistics*
Glen SATTEN, *Emory University*

Microbiome sequencing data are known to be biased; the measured taxa relative abundances can be systematically distorted from their true values at every step in the experimental/analysis workflow. If this bias is not accounted for, it can lead to spurious discoveries and invalid conclusions. Unfortunately, in order to measure bias it is necessary to have samples for which the true relative abundances are known, such as model or mock community samples. In this chapter, we propose a log-linear model for the biases observed when analyzing model communities data. Our model expands the recent work from McLaren, Willis and Callahan (MWC) [eLife, 8:e46923, 2019] that proposed a multiplicative bias structure for microbiome data. Our extension of the MWC model is general enough to allow testing of complex hypotheses, and readily handles situations in which samples have different number of bacteria present by design. An F-test with permutation-based hypothesis testing is proposed to assess statistical significance. We conduct simulations to show the validity and the power of our method, and also demonstrate the utility of our method through an analysis of a complex model communities dat=aset that allows us to directly test the multiplicative bias assumption of the MWC model. An R package implementing the proposed work is publicly available at https://github.com/zhaoni153/MicroBias.

212. Adaptation Development Strategy for Oncology Biomarker-Driven Registration Trial

[IS 38, (page 18)]

Xin ZHAO, *Oncolgoy Statistics,Janssen Pharceuticals*
Sudhakar RAO, *Janssen Pharceuticals*

Scientific knowledge and external data on biomarker mutations and associated treatment effect continues to evolve when an oncology program started the journey of planning and executing the phase 3 registration trial. Strategy is developed and adapted to overcome multiple challenges with evolving biomarker knowledge and regulatory environment, including initiating the registration trial with an innovative and flexible trial design framework to maximize the probability of success for the program with multi-level risk mitigation strategy, and the modification of the analysis strategy during the conduct of the registration trial to adapt to new regulatory environment based on emerging competitive landscape intelligence and extensive statistical modeling and simulations.

The innovative strategies developed in this journey can be utilized in a broader setting where a biomarker-driven clinical development is challenged with multi-level uncertainties.

213. Using Household and Retail Scanner Data to Inform Food and Nutrition Policy

[IS 13, (page 8)]

Chen ZHEN, *Agricultural and Applied Economics,University of Georgia*
Lan MU, *University of Georgia*
Gauri DATTA, *University of Georgia*
Chandra DHAKAL, *University of Georgia*

Tracking time series price variation is of great importance to governments, industries, researchers and consumers for inflation surveillance, and cost-of-living (COL) adjustments to salary and wage, public safety net program payments, and other income payments. Of equal importance is the within-country comparison of prices across locations. However, the latter is done less often because of a previous lack of proper data. The economic significance of spatial price difference has been demonstrated for federal tax distortion (Albouy 2009), poverty ranking (Jolliffe 2006), and purchasing power of fruit and vegetable voucher under the Women, Infants, and Children (WIC) program (Çakir et al. 2018) and food benefit under the Supplemental Nutrition Assistance Program (SNAP) (Zhen et al. 2018a). A COL index the type economists are most interested in is an index that tracks changes in the cost of maintaining a fixed standard of living. A panel index tracks temporal and spatial cost variations simultaneously. The objective of this project is to develop a public-use panel price database at the USDA food code level for small areas in the contiguous United States over the 20082017 period. To this end, we leverage big data on sales and purchases of 450,000 food barcodes at over 40,000 food retailers, state-of-the-art scanner data-based panel price indexes, and advanced GIS tools. The price database will be positioned to support a broad spectrum of research on food and nutrition policy, food market and price, and consumer demand. As a public good, this database will be a valuable addition to, but not a duplicate of, existing food price statistics reported by the Bureau of Labor Statistics, Bureau of Economic Analysis, Economic Research Service, and Council for Community and Economic Research.

214. Fast Approximation of Shapley Values

[IS 62, (page 28)]

Wei ZHENG, *Business Analytics and Statistics,university of tennessee*
liuqing YANG, *Nankai University*
Yongdao ZHOU, *Nankai University*
Haoda FU, *Eli Lilly and Company*
Minqian LIU, *Nankai University*

The Shapley value has been widely used in game theory, local model explanation and sensitivity analysis. However its calculation is an NP-complete problem. Specifically, calculating a d-player Shapley value requires evaluation of $d!$ or 2^d marginal contribution values, each associated with a permutations of the d players. Hence it becomes infeasible to calculate the Shapley value when d is too large. A common remedy is to take a random sample of the permutations to surrogate for the complete list of permutations. We find an advanced sampling scheme can be designed to yield much more accurate estimation of the Shapley value than the simple random sampling (SRS). Our sampling scheme is based on combinatorial structures in the field of design of experiment (DOE), particularly the order-of-addition experimental designs for the study of how the orderings of the components would affect the output. We show that the obtained estimates are unbiased and consistent, sometimes even deterministically recover the original Shapley value based on the full permutation. Both theoretical and simulations results show that our DOE based sampling scheme outperforms SRS.

215. A framework of Bayesian optimal phase II (BOP2) clinical trial design

[IS 46, (page 21)]

Heng ZHOU, *Biostatistics and Research Decision Sciences,Merck & Co., Inc*
Ying YUAN, *MD Anderson Cancer Center*
Linda SUN, *Merck & Co., Inc*
Cong CHEN, *Merck & Co., Inc*

We proposed a framework of Bayesian optimal phase II (BOP2) design for clinical trials with various simple and complex endpoints. The BOP2 design provides flexible interim go/no-go decisions which are made by comparing a set of posterior probabilities of the events of interest with adaptive probability cutoff. The BOP2 design can maximize the power with optimizing the interim probability cutoffs while explicitly control the type I error rate, thereby bridging the gap between Bayesian and frequentist designs. The BOP2 design is easy to implement with interim stopping boundaries available prior to the onset of the trial for most cases.

Directory

AMINI, Arash
Statistics, UCLA
aaamini@ucla.edu
Speaker: IS 31, p. 15, § 1, p. 35

ANDERES, Ethan
Department of Statistics, University of California at Davis
ethananderes@gmail.com
Speaker: IS 9, p. 6, § 2, p. 35

ANDERSON, Keaven
Methodology Research, Merck & Co., Inc.
keaven_anderson@merck.com
Speaker: IS 44, p. 20, § 3, p. 36

ARORA, Vipin
Eli Lilly and Company
varora@lilly.com
Chair and organizer: IS 8, p. 5

BALADANDAYUTHAPANI, Veera
Biostatistics, University of Michigan
veerab@umich.edu
Speaker: IS 25, p. 12, § 4, p. 36

BANERJEE, Kalins
Department of Public Health Sciences, Pennsylvania State University
email.kalins@gmail.com
Speaker: Student Paper Competition 2, p. 19, § 5, p. 36

BANERJEE, Mousumi
Biostatistics, University of Michigan
mousumib@umich.edu
Speaker: Special Invited Session 2, p. 7, § 6, p. 36

BANERJEE, Swarnali
Department of Mathematics and Statistics, Loyola University Chicago
swarnali009@gmail.com
Chair and organizer: IS 39, p. 18,
Speaker: IS 39, p. 18, § 7, p. 37

BANERJEE, Trambak
Analytics, Information and Operations Management, University of Kansas
trambakbanerjee@gmail.com
Chair: IS 53, p. 24,
Speaker: IS 53, p. 24, § 8, p. 37

BASU, Cynthia
Early Clinical Development, Pfizer Inc.
cynthia.basu@pfizer.com
Speaker: IS 61, p. 27, § 9, p. 38

BASU, Sanjib
School of Public Health, University of Illinois Chicago
sbasu@uic.edu
Chair: Conference Inaugaration, p. 3,
Chair: IS 29, p. 14

BASU, Saonli
Division of Biostatistics, University of Minnesota
saonli@umn.edu
Chair: Plenary Lecture 1, p. 3,
Speaker: Conference Inaugaration, p. 3,
Chair and organizer: IS 48, p. 22,
Speaker: IS 37, p. 17, § 10, p. 38

BASU, Sumanta
Statistics and Data Science, Cornell Uniersity
sumbose@cornell.edu
Speaker: IS 11, p. 6, § 11, p. 38

BASU, Sumanta
Statistics and Data Science, Cornell University
sumbose@cornell.edu
Organizer: IS 11, p. 6,
Chair and organizer: IS 59, p. 27

BERG, Emily
Statistics, Iowa State University
emilyb@iastate.edu
Speaker: IS 10, p. 6, § 12, p. 38

BHATTACHARYA, Anirban
Statistics, Texas A&M University
anirbanb@stat.tamu.edu
Speaker: IS 64, p. 28, § 13, p. 39

BHATTACHARYA, Bhaswar
Statistics, University of Pennsylvania
bhaswar.bhattacharya@gmail.com
Chair: IS 17, p. 9,
Speaker: IS 17, p. 9, § 14, p. 39

BHATTACHARYA, Rianka
Abbvie Inc.
riankabh@gmail.com
Chair and organizer: IS 27, p. 13

BHATTACHARYA, Sudipta
SQS, Biostatistics, Takeda
sudipta.bhattacharya@takeda.com
Speaker: IS 27, p. 13, § 15, p. 39

BHATTACHARYYA, Rupam
Biostatistics, University of Michigan
rupamb@umich.edu
Speaker: IS 23, p. 11, § 16, p. 40

BHATTACHARYYA, Sharmodeep
Oregon State University
bhattash@science.oregonstate.edu

Organizer: IS 54, p. 24

BISWAS, Swati
Mathematical Sciences, University of Texas at Dallas
swati.biswas@utdallas.edu
Chair and organizer: IS 37, p. 17,
Speaker: IS 48, p. 22, § 17, p. 40

BRETZ, Frank
Statistical Methodology, Novartis Pharma AG
frank.bretz@novartis.com
Speaker: Special Invited Session 4, p. 25, § 18, p. 41

CAO, Sky
Stanford Statistics, Stanfod University
skycao@stanford.edu
Speaker: Student Paper Competition 1, p. 15, § 19, p. 41

CASTRUCCIO, Stefano
Stefano Castruccio, University of Notre Dame
scastruc@nd.edu
Speaker: IS 6, p. 5, § 20, p. 41

CHAKRABORTY, Nilanjana
Statistics, University of Florida
nchakraborty@ufl.edu
Speaker: Student Paper Competition 1, p. 15, § 21, p. 41

CHAKRABORTY, Sounak
Statistics, University of Missouri-Columbia
chakrabortys@missouri.edu
Speaker: IS 42, p. 19, § 22, p. 42

CHAKRAVARTTY, Arunava
Biostatistics/Novartis, Novartis
archnova@gmail.com
Speaker: IS 61, p. 27, § 23, p. 42

CHAKRAVARTY, Aloka
Immediate Office of the Commissioner, FDA,US Food and Drug Administration
aloka.chakravarty@fda.hhs.gov
Speaker: Special Invited Session 4, p. 25, § 24, p. 42

CHAPPELL, Richard
U Wisconsin Dept. of Biostatistics and Medical Informatics,University of Wisconsin
chappell@stat.wisc.edu
Speaker: IS 55, p. 25, § 25, p. 43

CHATTERJEE, Ansu
School of Statistics, University of Minnesota
chatterjee@stat.umn.edu
Speaker: Conference Inaugaration, p. 3,
Speaker: IS 34, p. 16, § 26, p. 43

CHATTERJEE, Arkendu
Global Biometrics and Data Science, BMS,Associate Director, BMS
arkendu@gmail.com
Speaker: IS 27, p. 13, § 27, p. 43

CHATTERJEE, Nilanjan
615 N WOLFE ST, Suite E3527,Johns Hopkins University
nilanjan10c@gmail.com
Speaker: Plenary Lecture 1, p. 3, § 28, p. 43

CHATTERJEE, Sabyasachi
Statistics,University of Illinois at Urbana-Champaign
sc1706@illinois.edu
Chair and organizer: IS 47, p. 21,
Speaker: IS 17, p. 9, § 29, p. 44

CHATTERJEE, Shirshendu
City University of New York
shirshendu@ccny.cuny.edu
Organizer: IS 31, p. 14

CHAUDHURI, Sanjay
National University of Singapore
stasc@nus.edu.sg
Chair: Special Invited Session 1, p. 7,
Chair: IS 13, p. 8,
Chair: IS 45, p. 21

CHEN, Lin
Department of Public Health Sciences,University of Chicago
lchen@health.bsd.uchicago.edu
Speaker: IS 48, p. 22, § 30, p. 44

CHEN, Xiaotian
Data and Statistical Science, AbbVie
justinchen87@gmail.com
Speaker: IS 20, p. 10, § 31, p. 44

CHEN, Yuguo
Department of Statistics,University of Illinois at Urbana-Champaign
yuguo@illinois.edu
Speaker: IS 36, p. 17, § 32, p. 45

CHIB, Siddhartha
Olin Business School,Washington University in Saint Louis
chib@wustl.edu
Speaker: Special Invited Session 1, p. 7, § 33, p. 45

CHOI, David
Heinz College,Carnegie Mellon University
dave.s.choi@gmail.com
Speaker: IS 54, p. 24, § 34, p. 45

CHOWDHURY, SHRABANTI
Genetics and Genomic Sciences,Icahn school of Medicine at Mount Sinai
shrabanti.chowdhury@mssm.edu
Speaker: IS 18, p. 10, § 35, p. 45

CHUNG, Hee Cheol
Department of Statistics,Texas A&M University
heecheolchung31@gmail.com
Speaker: IS 13, p. 8, § 36, p. 46

CICCONETTI, Greg
Teri Anderson,AbbVie
gcicconetti@gmail.com
Speaker: IS 2, p. 3, § 37, p. 46

COOMBES, Brandon
Department of Quantitative Health Sciences, Mayo Clinic
coombes.brandon@mayo.edu
Chair and organizer: IS 65, p. 29,
Discussant: IS 65, p. 29,
Speaker: IS 65, p. 29, § 38, p. 47

CURSIO, John
Public Health Sciences,University of Chicago
jcursio@health.bsd.uchicago.edu
Speaker: IS 60, p. 27, § 39, p. 47

DAS, Sayan
Department of Mathematics,Columbia University
sd3225@columbia.edu
Speaker: Student Paper Competition 1, p. 15, § 40, p. 47

DATTA, Abhirup
Johns Hopkins University
abhidatta@jhu.edu
Chair: IS 11, p. 6,
Chair and organizer: IS 66, p. 29

DATTA, Gauri Sankar
Department of Statistics,University of Georgia/US Census Bureau
gauri@stat.uga.edu
Organizer: IS 13, p. 8,
Organizer: IS 45, p. 21,
Chair: IS 50, p. 23,
Speaker: IS 45, p. 21, § 41, p. 48

DATTA, Susmita
University of Florida
susmita.datta@ufl.edu
Chair: Plenary Lecture 3, p. 25

DAW, Ranadeep
University of Missouri, Department of Statistics,University of Missouri
rd2nr@mail.missouri.edu
Speaker: IS 45, p. 21, § 42, p. 48

DESAI, Neel M.
Rice University
nd26@rice.edu
Speaker: Student Paper Competition 2, p. 19, § 43, p. 48

DEY, Dipak
Department of Statistics,University of Connecticut
dey.dipak@gmail.com
Organizer: IS 58, p. 26,
Speaker: IS 29, p. 14, § 44, p. 49

DEY, Tanujit
Center for Surgery and Public Health, BWH,Harvard Medical School
tanujit.dey@gmail.com
Chair and organizer: IS 21, p. 11,
Organizer: IS 42, p. 19,
Speaker: IS 42, p. 19, § 45, p. 49

DUTTA, Somak
Statistics,Iowa State University
somakd@iastate.edu
Speaker: IS 66, p. 29, § 46, p. 49

FANG, Yixin
Data and Statistical Sciences,AbbVie
yixin.fang@abbvie.com
Speaker: IS 14, p. 8, § 47, p. 49

FENG, Dai
GMA Statistics, DSS, AbbVie,AbbVie
dai.feng@abbvie.com
Speaker: IS 30, p. 14, § 48, p. 49

FERREIRA, Marco
Marco Ferreira,Department of Statistics, Virginia Tech
marf@vt.edu
Speaker: IS 51, p. 23, § 49, p. 50

FITHIAN, Will
Statistics, UC Berkeley,UC Berkeley Statistics
wfithian@gmail.com
Speaker: IS 16, p. 9, § 50, p. 50

FU, Haoda
Eli Lilly and Company
fu_haoda@lilly.com
Organizer: IS 7, p. 5,
Speaker: IS 14, p. 8, § 51, p. 50

GAMALO, Margaret
GBDM-Inflammation and Immunology,Pfizer
meg.gamalo@yahoo.com
Speaker: IS 32, p. 16, § 52, p. 50

GHOSAL, Nairita
Merck & Co., INC.
nairita.ghosal89@gmail.com
Speaker: IS 52, p. 23, § 53, p. 51

GHOSAL, Nairita
Merck & Co., INC.
nairita.ghosal89@gmail.com
Chair and organizer: IS 52, p. 23

GHOSAL, Soutik
NICHD/DIPHR
soutik.ghosal@nih.gov
Chair: IS 18, p. 9

GHOSH, Joyee
Statistics and Actuarial Science,The University of Iowa
joyee123in@gmail.com
Chair and organizer: IS 35, p. 17,
Chair and organizer: IS 51, p. 23,
Speaker: IS 35, p. 17, § 54, p. 51

GHOSH, Pranab
Biostatistics,Pfizer Inc.
Prnbmth@gmail.com
Speaker: IS 2, p. 4, § 55, p. 51

GHOSH, Sujit
Department of Statistics,North Carolina State University
sujit.ghosh@ncsu.edu
Chair: Special Invited Session 3, p. 25,
Organizer: IS 57, p. 26,
Speaker: IS 44, p. 20, § 56, p. 52

GIESSING, Alexander
Department of Operations Research and Financial Engineering,Princeton University
giessing@princeton.edu
Speaker: IS 40, p. 18, § 57, p. 52

GILES, Wayne H.
Dean, School of Public Health, University of Illinois Chicago
wgiles@uic.edu
Speaker: Conference Inaugaration, p. 3

GUAN, Yawen
Statistics,University of Nebraska - Lincoln
yguan12@unl.edu
Speaker: IS 6, p. 5, § 58, p. 52

GUHANIYOGI, Rajarshi
Statistics, UC Santa Cruz
rguhaniy@ucsc.edu
Speaker: IS 6, p. 5, § 59, p. 52

GUNTUBOYINA, Adityanand
Statistics,University of California Berkeley
aditya@stat.berkeley.edu
Chair and organizer: IS 16, p. 9,
Speaker: IS 5, p. 4, § 60, p. 53

GUO, Wenge
Mathematical Sciences,New Jersey Institute of Technology
wenge.guo@njit.edu
Speaker: IS 43, p. 20, § 61, p. 53

GUPTA, Suyash
Statistics,Ph.D. student, Statistics, Stanford University
suyash28@stanford.edu
Speaker: Student Paper Competition 1, p. 15, § 62, p. 53

HALDER, Aritra
gxk9jg@virginia.edu
Biocomplexity Institute, University of Virginia
Chair: IS 58, p. 26

HEDEKER, Donald
University of Chicago
hedeker@uchicago.edu
Chair and organizer: IS 60, p. 27

HENDRICKSON, Barbara
Pharmacovigilance and Patient Safety, AbbVie,AbbVie
bhendric@bsd.uchicago.edu
Speaker: IS 26, p. 13, § 63, p. 54

HOOKER, Giles
Statistics and Data Science,Cornel University
giles.hooker@cornell.edu
Speaker: IS 59, p. 27, § 64, p. 54

HUANG, Erya
Clinical Statistics,Bayer U.S. LLC
erya.huang@bayer.com
Speaker: IS 26, p. 13, § 65, p. 54

HUANG, Xin
Discovery and Exploratory Statistics (DIVES), Data & Statistical Sciences,AbbVie Inc.
xin.huang@abbvie.com
Organizer: IS 14, p. 8,
Speaker: IS 14, p. 8, § 66, p. 54

HYUN, Noorie
Division of Biostatistics,Medical College of Wisconsin
nhyun@mcw.edu
Speaker: IS 24, p. 12, § 67, p. 55

IPE, David
Data and Statistical Science, AbbVie
david.ipe@abbvie.com
Speaker: IS 20, p. 10, § 68, p. 55

ISHII, Aki
Department of Information Sciences,Tokyo University of Science
aki.0321tk@gmail.com
Speaker: IS 28, p. 14, § 69, p. 55

JANKAR, Jeevan
Department of Statistics,University of Georgia, Athens
jeevan.jankar@uga.edu
Speaker: IS 56, p. 26, § 70, p. 55

JEMIAI, Yannis
Cytel,Cytel
yannis.jemiai@cytel.com
Speaker: IS 19, p. 10, § 71, p. 56

JEWELL, Cathleen
Clinical Analytics,AbbVie
cathleen.jewell@abbvie.com
Speaker: IS 8, p. 6, § 72, p. 56

JIANG, Ruochen
Statistics, University of California, Los Angeles
ruochenj@gmail.com
Speaker: Student Paper Competition 2, p. 20, § 73, p. 56

JONES, David
Department of Statistics, Texas A&M University
djones@stat.tamu.edu
Speaker: IS 22, p. 11, § 74, p. 56

JOO, Mingyu (Max)
Marketing, UC Riverside
mingyu.joo@ucr.edu
Speaker: IS 50, p. 23, § 75, p. 57

JOSEPH, Roshan
School of Industrial and Systems Engineering, Georgia Institute of Technology
roshan@gatech.edu
Speaker: Special Invited Session 3, p. 25, § 76, p. 57

KAIZER, Alexander
Biostatistics and Informatics, University of Colorado-Anschutz Medical Campus
alex.kaizer@cuanschutz.edu
Speaker: IS 21, p. 11, § 77, p. 57

KANG, Jian
Jian Kang, University of Michigan
jiankang@umich.edu
Speaker: IS 25, p. 12, § 78, p. 57

KAR, Wreetabrata
Purdue University, Purdue University-Main Campus (West Lafayette, IN)
wkar@purdue.edu
Speaker: IS 53, p. 24, § 79, p. 58

KARMAKAR, Bikram
Statistics, University of Florida
bkarmakar@ufl.edu
Speaker: IS 4, p. 4, § 80, p. 58

KARNOUB, Maha
Maha Karnoub, Daiichi Sankyo - Translational Medicine Biostatistics
mkarnoub@dsi.com
Speaker: IS 38, p. 18, § 81, p. 58

KASHYAP, Vinay
High-Energy Astrophysics Division, Center for Astrophysics – Harvard & Smithsonian
vkashyap@cfa.harvard.edu
Speaker: IS 22, p. 11, § 82, p. 59

KAUR, Amarjot
Merck & Co., Inc.
amarjot_kaur@merck.com
Chair: Plenary Lecture 2, p. 13

KHARE, Kshitij
Department of Statistics, University of Florida
kdkhare@stat.ufl.edu
Chair and organizer: IS 64, p. 28,
Speaker: IS 64, p. 29, § 83, p. 59

KIM, Soyoung
Division of Biostatistics, Medical College of Wisconsin
skim@mcw.edu
Speaker: IS 24, p. 12, § 84, p. 59

KLUSOWSKI, Jason
Operations Research and Financial Engineering, Princeton University
jason.klusowski@princeton.edu
Speaker: IS 59, p. 27, § 85, p. 60

KOLAR, Mladen
The University of Chicago Booth School of Business, The University of Chicago Booth School of Business
mkolar@chicagobooth.edu
Speaker: IS 28, p. 13, § 86, p. 60

KONG, Xiaoli
Department of Mathematics and Statistics, Loyola University Chicago
xkong1@luc.edu
Speaker: IS 39, p. 18, § 88, p. 60

KUCHIBHOTLA, Arun
Department of Statistics and Data Science, Carnegie Mellon University
karun3kumar@gmail.com
Speaker: IS 5, p. 4, § 89, p. 61

KULKARNI, Hrishikesh
Alexion Pharmaceuticals
hrishikesh.kulkarni@alexion.com
Chair and organizer: IS 19, p. 10,
Chair and organizer: IS 26, p. 13

KUNDU, Madan
Daiichi Sankyo Inc
mkundu@dsi.com
Chair and organizer: IS 2, p. 3,
Speaker: IS 49, p. 22, § 90, p. 61

KUNDU, Suprateek
Biostatistics, Emory University
Suprateek.kundu@emory.edu
Speaker: IS 42, p. 19, § 91, p. 61

LAKSHMINARAYAN, Choudur
Teradata Labs, Teradata Labs
choudur.lakshminarayan@utexas.edu
Speaker: IS 41, p. 19, § 92, p. 62

LAN, Zhou
Yale School of Medicine, Yale School of Medicine
zhou.lan@yale.edu
Speaker: IS 66, p. 29, § 93, p. 62

LAUD, Purushottam
Medical College of Wisconsin
laud@mcw.edu
Chair and organizer: IS 15, p. 8,
Chair: IS 23, p. 11,
Speaker: IS 58, p. 26, § 94, p. 62

LEE, Jae
Biomarker Group, Global Biometric and Data Sciences, Bristol Myers Squibb
jae.lee@bms.com
Speaker: IS 38, p. 18, § 95, p. 62

LEON, Larry
Biostatistics, Bristol-Myers Squibb
larry.leon.05@post.harvard.edu
Speaker: IS 61, p. 28, § 96, p. 62

LI, Dingzhou (Dean)
Drug Safety Statistics, Pfizer, Drug Safety Statistics, Pfizer
dingzhou.li@pfizer.com
Speaker: IS 52, p. 23, § 97, p. 63

LI, Tianxi
Department of Statistics, University of Virginia
tianxili@virginia.edu
Speaker: IS 54, p. 24, § 98, p. 63

LI, Xiao
Biostatistics, Medical College of Wisconsin
xiaoli@mcw.edu
Speaker: IS 15, p. 9, § 99, p. 63

LI, Xiaodong
UC Davis/Statistics, UC Davis
xdgli@ucdavis.edu
Speaker: IS 9, p. 6, § 100, p. 64

LI, Xinran
Department of Statistics, University of Illinois at Urbana-Champaign
xinranli@illinois.edu
Speaker: IS 56, p. 26, § 101, p. 64

LIANG, Feng
Statistics Department, UIUC, University of Illinois at Urbana-Champaign
liangf@illinois.edu
Speaker: IS 63, p. 28, § 102, p. 64

LIN, Ruitao
Ruitao Lin, The University of Texas MD Anderson Cancer Center
ruitaolin@gmail.com
Speaker: IS 46, p. 21, § 103, p. 65

LIN, Shili
Statistics, Ohio State University
shili@stat.osu.edu
Speaker: IS 37, p. 17, § 104, p. 65

LIN, Xihong
Departments of Biostatistics and Department of Statistics, Harvard University and Broad Institute
xlin@hsph.harvard.edu
Speaker: Plenary Lecture 3, p. 25, § 105, p. 65

LIPKOVICH, Ilya
Real world analytics, Eli Lilly and Company
ilya.lipkovich@lilly.com
Speaker: IS 1, p. 3, § 106, p. 66

LIU, Alex
Bayer
alex.liu1@bayer.com
Chair: IS 46, p. 21

LIU, Feng
Oncology Biometrics, AstraZeneca
fenglpku2000@gmail.com
Organizer: IS 55, p. 25,
Speaker: IS 55, p. 25, § 107, p. 66

LIU, Rong
Bristol Myers Squibb
rong.liu1@bms.com
Chair and organizer: IS 38, p. 18,
Organizer: IS 46, p. 21

LU, Qiongshi
Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison
Qiongshi.lu@gmail.com
Speaker: IS 12, p. 7, § 108, p. 66

MA, Ping
Statistics, University of Georgia
pingma@uga.edu
Speaker: IS 3, p. 4, § 109, p. 67

MA, Qianheng
Department of Public Health Sciences, University of Chicago
qma@uchicago.edu
Speaker: IS 60, p. 27, § 110, p. 67

MADRID PADILLA, OSCAR HERNAN
Statistics, University of California, Los Angeles
oscar.madrid@stat.ucla.edu
Speaker: IS 47, p. 22, § 111, p. 67

MAITY, Arnab
Pfizer
Arnab.Maity@pfizer.com
Chair and organizer: IS 32, p. 15,
Chair and organizer: IS 49, p. 22,
Chair and organizer: IS 61, p. 27

MAJUMDAR, Antara
Statistical Innovations Group, Medidata Acorn AI
amajumdar@mdsol.com
Speaker: IS 30, p. 14, § 112, p. 68

MAJUMDAR, Dibyen
Math., Stat. and Comp., Sci, UIC, University of Illinois at Chicago
dibyen@uic.edu
Chair: IS 3, p. 4,
Speaker: IS 50, p. 23, § 113, p. 68

MAJUMDAR, Subhabrata
Data science and AI Research, University of Minnesota
zoom.subha@gmail.com
Speaker: IS 41, p. 19, § 114, p. 68

MAK, Simon
Statistical Science, Duke University
sm769@duke.edu
Speaker: IS 57, p. 26, § 115, p. 69

MALLICK, Bani
Statistics, Texas A&M
bmallick@stat.tamu.edu
Speaker: Special Invited Session 3, p. 25, § 116, p. 69

MALLICK, Himel
Biostatistics and Research Decision Sciences, Merck Research Laboratories
himel.mallick@merck.com
Organizer: IS 18, p. 9

MALLICK, Himel
Biostatistics and Research Decision Sciences, Merck Research Laboratories
himel.stat.iitk@gmail.com
Speaker: IS 18, p. 10, § 117, p. 69

MANDAL, Abhijit
Department of Mathematical Sciences, University of Texas at El Paso
abhijit.mandal@wayne.edu
Chair: IS 34, p. 16,
Speaker: IS 34, p. 16, § 118, p. 70

MANDAL, Abhyuday
University of Georgia
amandal@stat.uga.edu
Organizer: IS 3, p. 4,
Organizer: IS 50, p. 23,
Organizer: IS 56, p. 26,
Organizer: IS 62, p. 28,
Speaker: IS 13, p. 8, § 119, p. 70

MAO, Lu
Biostatistics and Medical Informatics, University of Wisconsin-Madison
lmao@biostat.wisc.edu
Speaker: IS 24, p. 12, § 120, p. 70

MATTHEWS, Gregory
Mathematics and Statistics, Loyola University Chicago
gmatthews1@luc.edu
Speaker: IS 39, p. 18, § 121, p. 71

MCCULLOCH, Robert
School of Mathematical and Statistical Sciences, Arizona State
robert.e.mcculloch@gmail.com
Speaker: IS 15, p. 8, § 122, p. 71

MENTCH, Lucas
Department of Statistics, University of Pittsburgh
lkm31@pitt.edu
Speaker: IS 59, p. 27, § 123, p. 71

MONDAL, Anirban
Department of Mathematics, Applied Mathematics, and Statistics, Case Western Reserve University
axm912@case.edu
Organizer: IS 34, p. 16,
Chair and organizer: IS 41, p. 19,
Speaker: IS 34, p. 16, § 124, p. 72

MONDAL, Debashis
Oregon State University
debashis@stat.oregonstate.edu
Chair: IS 31, p. 14,
Chair: IS 54, p. 24

MOTTA, Giovanni
Statistics,Texas A&M, Department of Statistics
g.motta@stat.tamu.edu
Speaker: IS 22, p. 11, § 125, p. 72

MUELLER, Peter
Statistics & Data Sc,UT Austin
pmueller@math.utexas.edu
Speaker: IS 58, p. 26, § 126, p. 72

MUKHERJEE, Gourab
Department of Data Sciences & Operations,University of Southern California
gourab@usc.edu
Organizer: IS 17, p. 9,
Organizer: IS 53, p. 24,
Speaker: IS 9, p. 6, § 127, p. 73

MUKHERJEE, Soumendu Sundar
Interdisciplinary Statistical Research Unit (ISRU),Indian Statistical Institute, Kolkata
soumendu041@gmail.com
Speaker: IS 31, p. 15, § 128, p. 73

MUKHERJEE, Sumit
Sumit Mukherjee,Columbia University
sm3949@columbia.edu
Speaker: IS 17, p. 9, § 129, p. 73

MUKHOPADHYAY, Pabak
Daiichi-Sankyo Inc
PMUKHOPADHY@DSI.COM
Chair and organizer: IS 30, p. 14,
Discussant: IS 43, p. 20

MUKHOPADHYAY, Pralay
Department of Biometrics,Otsuka America Pharmaceuticals Inc.
pralay.mukhopadhyay@otsuka-us.com

Chair: Special Invited Session 2, p. 7,
Chair and organizer: IS 44, p. 20,
Chair: IS 55, p. 25,
Speaker: IS 44, p. 21, § 130, p. 73

MUKHOPADHYAY, Saurabh
Data and Statistical Science, AbbVie
stat.mukherjee@gmail.com
Discussant: IS 20, p. 10,
Organizer: IS 20, p. 10

MUKHOTI, Sujay
Operations Management and Quantitative Techniques,Indian Institute of Management Indore
sujay.mukhoti@gmail.com
Speaker: IS 4, p. 4, § 131, p. 74

MUNSAKA, Melvin
Statistical Sciences,AbbVie
melvin.munsaka@abbvie.com
Speaker: IS 7, p. 5, § 132, p. 74

NARISETTY, Naveen Naidu
Statistics,University of Illinois at Urbana-Champaign
naveen@illinois.edu
Organizer: IS 40, p. 18,
Chair and organizer: IS 63, p. 28,
Speaker: IS 40, p. 18, § 133, p. 74

NATARAJAN, Kannan
Global Head of Biometrics and Data Management,Pfizer
Kannan.Natarajan@pfizer.com
Speaker: Plenary Lecture 2, p. 13, § 134, p. 75

NEWTON, Michael
Department of Statistics and Department of Biostatistics and Medical Informatics,University of Wisconsin-Madison
newton@biostat.wisc.edu
Speaker: Special Invited Session 1, p. 7, § 135, p. 75

PADDOCK, Susan
Statistics and Methodology,NORC at the University of Chicago
paddock-susan@norc.org
Speaker: Special Invited Session 2, p. 7, § 136, p. 75

PAJDA-DE LA O, Jennifer
Mathematics, Statistics, and Computer Science,University of Illinois at Chicago
jpajda2@uic.edu
Organizer: IS 10, p. 6,
Speaker: IS 10, p. 6, § 137, p. 76

PALMER, Jeff
Early Clinical Development Statistics,Pfizer, Inc.
jeffrey.p.palmer@pfizer.com
Speaker: IS 19, p. 10, § 138, p. 76

PATI, Debdeep
Statistics,Texas A&M University
debdeep@stat.tamu.edu
Speaker: IS 64, p. 29, § 139, p. 76

PATIL, Sujata
Cleveland Clinic
PATILS2@ccf.org
Chair: Student Paper Competition 2, p. 19

PATRA, Rohit
Department of Statistics,University of Florida
rohitpatra@ufl.edu
Speaker: IS 5, p. 5, § 140, p. 76

PAUL, Debashis
University of California, Davis
debpaul@ucdavis.edu
Chair and organizer: IS 9, p. 6

PAUL, Erina
Biostatistics and Research Decision Sciences,Merck & Co., Inc.
erina.paul633@gmail.com
Speaker: IS 27, p. 13, § 141, p. 77

PAUL, Subhadeep
Department of Statistics,The Ohio State University
paul.963@osu.edu
Speaker: IS 36, p. 17, § 142, p. 77

PRADHAN, Vivek
Statistics, ECD I&I,Pfizer Inc
vivek.pradhan@pfizer.com
Speaker: IS 49, p. 22, § 143, p. 77

PSIODA, Matthew
Department of Biostatistics,Department of Biostatistics
matt_psioda@unc.edu
Speaker: IS 21, p. 11, § 144, p. 78

PUELZ, David
Booth School of Business,University of Chicago
David.Puelz@chicagobooth.edu
Speaker: IS 56, p. 26, § 145, p. 78

QIN, Qian
School of Statistics,University of Minnesota
qqin@umn.edu
Speaker: IS 15, p. 8, § 146, p. 78

QU, Yongming
Department of Data and Analytics,Eli Lilly and Company
qu_yongming@lilly.com
Speaker: IS 1, p. 3, § 147, p. 79

RAHNAVARD, Ali
Biostatistics and Bioinformatics,George Washington University
rahnavard@gwu.edu
Speaker: IS 18, p. 10, § 148, p. 79

RAY, Debashree
Johns Hopkins University
dray@jhu.edu
Chair and organizer: IS 12, p. 7

RETTIGANTI, Mallikarjuna
Neuroscience,Eli Lilly and Company
rettiganti_mallikarjuna@lilly.com
Speaker: IS 8, p. 6, § 149, p. 79

ROY, Arkaprava
University of Florida
ark007@ufl.edu
Speaker: IS 4, p. 4, § 150, p. 79

ROY, Pourab
Office of Biostatistics,FDA
Pourab.Roy@fda.hhs.gov
Speaker: IS 30, p. 14, § 151, p. 80

ROY, Vivekananda
Department of Statistics,Iowa State University
vroy@iastate.edu
Speaker: IS 35, p. 17, § 152, p. 80

ROYCHOUDHURY, Satrajit
Pfizer Inc.
satrajit.roychoudhury@pfizer.com
Organizer: IS 43, p. 20

RUDRA, Pratyaydipta
Statistics,Oklahoma State University
prudra@okstate.edu
Speaker: IS 12, p. 7, § 153, p. 80

SABBAGHI, Arman
Department of Statistics,Purdue University
sabbaghi@purdue.edu
Speaker: IS 3, p. 4, § 154, p. 81

SADHANALA, Veeranjaneyulu
Booth School of Business,University of Chicago
veeranjaneyulus@gmail.com
Speaker: IS 47, p. 22, § 155, p. 81

SAFIKHANI, Abolfazl
University of Florida,University of Florida
a.safikhani@ufl.edu
Speaker: IS 11, p. 7, § 156, p. 81

SAHOO, Indranil
Statistical Sciences & Operations Research,Virginia Commonwealth University
sahooi@vcu.edu
Chair and organizer: IS 6, p. 5,
Speaker: IS 66, p. 30, § 157, p. 81

SALIL, Koner
North Carolina State University
skoner@ncsu.edu
Speaker: Student Paper Competition 1, p. 15, § 87, p. 60

SARKAR, Abhra
Statistics and Data Sciences,The University of Texas at Austin
abhra.stat@gmail.com
Speaker: IS 41, p. 19, § 158, p. 82

SARKAR, Purnamrita
Department of Statistics and Data Sciences,Asst. Prof.
purna.sarkar@austin.utexas.edu
Speaker: IS 31, p. 15, § 159, p. 82

SCHOLTENS, Denise
Department of Preventive Medicine - Biostatistics,Department of Preventive Medicine - Biostatistics
dscholtens@northwestern.edu
Speaker: IS 48, p. 22, § 160, p. 82

SEN, Ananda
Department of Biostatistics, University of Michigan
anandas@umich.edu
Organizer: IS 29, p. 14,
Speaker: IS 29, p. 14, § 161, p. 83

SEN, Bodhisattva
Department of Statistics,Columbia University
bodhi@stat.columbia.edu
Chair and organizer: IS 5, p. 4,
Chair: Student Paper Competition 1, p. 15,
Speaker: IS 16, p. 9, § 162, p. 83

SENGUPTA, Srijan
Statistics,North Carolina State University
ssengup2@ncsu.edu
Chair and organizer: IS 36, p. 17,
Speaker: IS 36, p. 17, § 163, p. 83

SETHURAMAN, Shanthi
Eli Lilly and Company
sethuraman_shanthi@lilly.com
Chair and organizer: IS 1, p. 3,
Chair: Special Invited Session 4, p. 25

SHARPNACK, James
Statistics Department,UC Davis
jsharpna@ucdavis.edu
Speaker: IS 47, p. 22, § 164, p. 84

SHIN, Sunyoung
University of Texas at Dallas
sunyoung.shin@utdallas.edu
Chair and organizer: IS 24, p. 12

SHOJAIE, Ali
Biostatistics,University of Washington
ali.shojaie@gmail.com
Speaker: IS 11, p. 7, § 165, p. 84

SI, Yajuan
Survey Research Center,University of Michigan
yajuan@umich.edu
Speaker: IS 35, p. 17, § 166, p. 84

SIDDANI, Satya Ravi
Abbvie Inc.
Chair: IS 43, p. 20

SINGH, Rakhi
Informatics and Analytics,University of North Carolina at Greensboro
r_singh5@uncg.edu
Speaker: IS 62, p. 28, § 167, p. 85

SINGH, Satya
Mathematics,Indian Institute of Technology Hyderabad
spsingh@iith.ac.in
Speaker: IS 62, p. 28, § 168, p. 85

SINHA, Samiran
Texas A&M University
sinha@stat.tamu.edu
Chair and organizer: IS 22, p. 11

SIVAGANESAN, Siva

Division of Statistics and Data Science, Department of Mathemtical Sciences,University of Cincinnati

sivagas@ucmail.uc.edu

Speaker: IS 58, p. 26, § 169, p. 85

SLOAN, Abigail

Biostatistics,Pfizer

abigail.sloan@pfizer.com

Speaker: IS 8, p. 5, § 170, p. 85

SPANBAUER, Charles

Biostatistics,University of Minnesota

spanb008@umn.edu

Speaker: IS 23, p. 12, § 171, p. 85

SPARAPANI, Rodney

Division of Biostatistics,Medical College of Wisconsin

rsparapa@mcw.edu

Organizer: IS 23, p. 11,

Speaker: IS 23, p. 12, § 172, p. 86

SRIRAM, Karthik

Indian Institute of Management Ahmedabad,India

karthiks@iima.ac.in

Chair and organizer: IS 4, p. 4

SRIVASTAVA, Sanvesh

Department of Statistics and Actuarial Science,The University of Iowa

sanvesh-srivastava@uiowa.edu

Speaker: IS 51, p. 23, § 173, p. 86

STUFKEN, John

Informatics and Analytics,University of North Carolina at Greensboro

jstufken@uncg.edu

Chair: IS 57, p. 26,

Chair: IS 62, p. 28,

Speaker: IS 50, p. 23, § 174, p. 86

SUN, Dongchu

Statistics,University of Nebraska-Lincoln

dsun9@unl.edu

Speaker: IS 45, p. 21, § 175, p. 87

SUN, Steven

Statistical Decision Science,Janssen R&D

ssun@its.jnj.com

Speaker: IS 49, p. 22, § 176, p. 87

SUN, Wenguang

University of Southern California,Session on Empirical Bayes Methodology

wenguans@marshall.usc.edu

Speaker: IS 16, p. 9, § 177, p. 87

SUN, Yan

Abbvie Inc.

Chair: IS 14, p. 8

SUSSMAN, Daniel

Mathematics and Statistics,Boston University

dpmcsuss@gmail.com

Speaker: IS 54, p. 24, § 178, p. 88

TADESSE, Mahlet

Department of Mathematics and Statistics,Georgetown University

mgt26@georgetown.edu

Speaker: IS 51, p. 23, § 179, p. 88

TAKEDA, Kentaro

Data Science,Astellas Pharma Global Development, Inc

kentaro.takeda@astellas.com

Speaker: IS 26, p. 13, § 180, p. 88

TAMHANE, Ajit

Ajit Tamhane,Northwestern University

ajit@iems.northwestern.edu

Speaker: IS 43, p. 20, § 181, p. 88

THOMAS, Neal

Pfizer, Groton CT USA,Pfizer

snthomas99@gmail.com

Speaker: IS 32, p. 16, § 182, p. 89

TIAN, Tian

Biostatistics,BeiGene

jta.1879@gmail.com

Speaker: IS 33, p. 16, § 183, p. 89

TURKMEN, Asuman

Department of Statistics,OHIO STATE UNIVERSITY

turkmen.2@osu.edu

Speaker: IS 37, p. 17, § 184, p. 89

VENKATRAMAN, Sara

Department of Statistics and Data Science,Cornell University, Department of Statistics and Data Science

skv24@cornell.edu

Speaker: Student Paper Competition 2, p. 20, § 185, p. 89

WAHAB, Hakeem

Statistics,Purdue University

aabdulw@purdue.edu

Speaker: IS 1, p. 3, § 186, p. 90

WANG, Hai Ying

Statistics,University of Connecticut

whygps@gmail.com

Speaker: IS 57, p. 26, § 187, p. 90

WANG, Hongwei

Global Medical Affairs Statistics,AbbVie

hongwei.wang@abbvie.com

Speaker: IS 7, p. 5, § 188, p. 90

WANG, Jingshu

Statistics,University of Chicago

jingshuw@uchicago.edu

Speaker: IS 28, p. 14, § 189, p. 91

WANG, Li

Data and Statistical Science, AbbVie

li.wang1@abbvie.com

Chair: IS 20, p. 10

WANG, Li

Data and Statistical Sciences, AbbVie

li.wang1@abbvie.com

Chair: IS 7, p. 5

WANG, Lin

Department of Statistics,George Washington University

linwang@gwu.edu

Speaker: IS 57, p. 26, § 190, p. 91

WANG, Ling

Worldwide Research, Development and Medical,Pfizer

ling.wang2@pfizer.com

Speaker: IS 32, p. 15, § 191, p. 91

WANG, Selena

Ohio State University

wang.10171@osu.edu

Speaker: Student Paper Competition 2, p. 20, § 192, p. 91

WANG, Zailong

AbbVie,AbbVie

zailong.wang@abbvie.com

Speaker: IS 19, p. 10, § 193, p. 92

WELCH, Whitney

Department of Preventive Medicine,Northwestern University Feinberg School of Medicine

whitney.welch@northwestern.edu

Speaker: IS 60, p. 27, § 194, p. 92

WU, Chong
Department of Statistics, Florida State University,Department of Statistics, Florida State University
cwu3@fsu.edu
Speaker: IS 65, p. 29, § 195, p. 93

XING, Yunzhao
Data and Statistical Science,AbbVie
yunzhao.xing@abbvie.com
Speaker: IS 7, p. 5, § 196, p. 93

YAN, Dongyan
Discovery & Development Statistics,Eli Lilly and Company
ydyhfutheaven@gmail.com
Speaker: IS 52, p. 23, § 197, p. 93

YANG, Jie
Department of Mathematics, Statistics, and Computer Science,University of Illinois at Chicago
jyang06@uic.edu
Chair: IS 56, p. 26,
Speaker: IS 3, p. 4, § 198, p. 94

YANG, Min
Department of Mathematics, Statistics, and Computer Science,University of Illinois at Chicago
myang2@uic.edu
Chair: IS 10, p. 6,
Chair and organizer: IS 33, p. 16,
Speaker: IS 33, p. 16, § 199, p. 94

YANG, Yun
Statistics,University of Illinois Urbana-Champaign
yy84@illinois.edu
Speaker: IS 40, p. 18, § 200, p. 94

YE, Jiabu
Oncology Biometrics AstraZeneca,AstraZeneca
jiabu.ye@merck.com
Speaker: IS 55, p. 25, § 201, p. 95

YUAN, Ying
Biostatistics,University of Texas MD Anderson Cancer Center
yyuan@mdanderson.org
Speaker: IS 33, p. 16, § 202, p. 95

ZABOR, Emily
Department of Quantitative Health Sciences, Cleveland Clinic
zabore2@ccf.org
Chair: IS 42, p. 19,
Speaker: IS 21, p. 11, § 203, p. 95

ZHAN, Tianyu
Data and Statistical Sciences, AbbVie Inc.,Data and Statistical Sciences, AbbVie Inc.
tianyu.zhan@abbvie.com
Speaker: IS 2, p. 3, § 204, p. 96

ZHANG, Anru
Statistics,University of Wisconsin-Madison / Duke University
anru.stat@gmail.com
Speaker: IS 63, p. 28, § 205, p. 96

ZHANG, Hongtao
Global Biometrics and Data Sciences,Bristol Myers Squibb
hongtao.zhang@bms.com
Speaker: IS 46, p. 21, § 206, p. 96

ZHANG, Lin
Division of Biostatistics, University of Minnesota
zhan4800@umn.edu
Chair and organizer: IS 25, p. 12,
Speaker: IS 25, p. 12, § 207, p. 97

ZHANG, Teng
Teng Zhang,University of Central Florida
teng.zhang@ucf.edu
Speaker: IS 47, p. 22, § 208, p. 97

ZHANG, Yuan
Statistics,The Ohio State University
yzhanghf@stat.osu.edu
Chair: IS 40, p. 18,
Speaker: IS 63, p. 28, § 209, p. 97

ZHAO, Dave
Statistics,University of Illinois at Urbana-Champaign
dave.zhao@gmail.com
Speaker: IS 53, p. 24, § 210, p. 97

ZHAO, Ni
Johns Hopkins University,assistant professor of Biostatistics
nzhao10@jhu.edu
Speaker: IS 12, p. 7, § 211, p. 98

ZHAO, Xin
Oncolgoy Statistics,Janssen Pharceuticals
xzhao121@gmail.com
Speaker: IS 38, p. 18, § 212, p. 98

ZHEN, Chen
Agricultural and Applied Economics,University of Georgia
czhen@uga.edu
Speaker: IS 13, p. 8, § 213, p. 98

ZHENG, Wei
Business Analytics and Statistics,university of tennessee
wzheng9@utk.edu
Speaker: IS 62, p. 28, § 214, p. 99

ZHONG, Ping-Shou
University of Illinois at Chicago
pszhong@uic.edu
Chair and organizer: IS 28, p. 13

ZHOU, Heng
Biostatistics and Research Decision Sciences,Merck & Co., Inc
heng.zhou@merck.com
Speaker: IS 46, p. 21, § 215, p. 99