The Western North American
Region of The International Biometric Society

2021 Annual Meeting

WNAR / IMS / JR

June 11 – 16, 2021

Program Book

Virtual Conference

Table of Contents

PRESIDENT'S WELCOME - JUNE 2021 3

WNAR OFFICERS 4

PRESIDENTIAL INVITED ADDRESS 6

SHORT COURSE 1 7

SHORT COURSE 2 8

SCIENTIFIC PROGRAM 9

ABSTRACTS 25

President's Welcome - June 2021

Welcome to the 2021 WNAR/IMS/JR virtual Conference! I would like to acknowledge that the lands and waters where WNAR members work and gather are the traditional territories of Indigenous peoples. We are committed to building respectful and collaborative relationships with these communities.

We have had a challenging year and a half navigating our way through the pandemic and I would like to thank those who helped make critical decisions especially Ying Lu (Past-President) and the Regional Committee members. I am looking forward to virtually come together as a community to meet you online. We have an exciting program with a diverse range of invited and contributed sessions. Our Student Paper Competition is overwhelmingly popular this year and we have two short courses on offer "Master Protocols: Tackling Complex Diseases (and COVID-19!) with Bayesian Adaptive Platform Trials" by Ben Saville, Anna McGlothlin, and Christina Saunders (Berry Consultants), and "RMST-based survival analysis methods for non-proportional hazards" by Lu Tian (Stanford University). I am also excited to increase the ecological statistics presence through the presentation by Rachel Fewster, Professor, Department of Statistics, University of Auckland, who will give the Presidential Invited Address entitled "How to count the things you didn't see: the magic and mystery of estimating population size".

Scientific Program Chair Yingqi Zhao and IMS Program Chair Lexin Li (UC Berkeley) have done an amazing job putting our program together including an array of invited sessions and student presentations. I extend my gratitude towards Audrey Hendricks who is running the WHOVA team behind-the-scenes. This year's conference is also a joint effort with the Japanese Region (JR), and we thank President Shigeyuki Matsui (Department of Biostatistics, Nagoya University Graduate School of Medicine) for partnering with WNAR to promote collaboration between our two regions. Many thanks to Megan Othus, Jessica Minnier (Secretaries), and Brandie Wagner (Treasurer) for guidance. Also special thanks to local arranger Jiaqi Huang (USFW) who has negotiated our postponement of the Alaska conference to 2023. I would also like to acknowledge the generous commitment of our Student Paper Competition Committee: Laura Saba (chair), Jarrett Barber, Cindy Feng, Camille Moore, Holly Steeves, Mourad Tighiouart, Julie Zhou. Thank you all for your hard work in making this conference a reality!

This is the first year WNAR has received industry sponsorship and I would like to acknowledge Beigene for their support. Finally, I would like to thank everyone for joining us online at this year's WNAR/IMS/JR conference. I hope that everyone learns something new, meets someone new, and catches up with old friends.

Laura Cowen
2021 WNAR President

WNAR Officers

of the Western North American Region of the International Biometric Society

Elected Officers and Representatives (Regional Committee)

Office	Year(s)	Officer	Affiliation
President	2021	Laura Cowen	University of Victoria
Secretary/Correspondent	2020-21	Megan Othus	Fred Hutchinson Cancer Research Center
Secretary- Elect	2021	Jessica Minnier	Oregon Health & Science University
Treasurer	2020-23	Brandie Wagner	University of Colorado Anschutz Medical Campus
Program Coordinator	2021-22	Elizabeth Jaurez-Colunga	University of Colorado Anschutz Medical Campus
President Elect	2019	Gary Chan	University of Washington
Past President	2019	Ying Lu	Stanford University

Representatives At-Large

2019-21	Katie Kerr	University of Washington
2020-22	Lisa Brown	Seattle Genetics
2020-22	Karen Messer	University of California San Diego
2021-23	Charlotte Gard	New Mexico State University
2021-23	Julia Palacios	Stanford University

IBS Council Representatives

2019-23	Elizabeth Jaurz-Colunga	University of Colorado Anschutz Medical Campus
2013-21	Antje Hoering	Cancer Research and Biostatistics
2017-21	Layla Parast	RAND

Appointments

RAB – Regional Advisory Board

Years	Chair	Members	Affiliation
2019-21		Jay Barber	Northern Arizona University
		Ann Lazar	University of California San Francisco
		Camille Moore	National Jewish Health
		Laura Saba	University of Colorado Anschutz Medical Campus
2020-22		Fang Chen	SAS
	2021	Joan Hu	Simon Fraser University
		David Rocke	University of California Davis
2021-23		Ruth Joy	Simon Fraser University
		Linglong Kong	University of Alberta
		Leslie New	Washington State University

2021-22	Jennifer McNichol (student rep	University of New Brunswick

Communications Officer

2021-22	Mieke Niederhausen	Oregon Health & Science University

WNAR Conference Local Organizers

2021 Program Chair	Yingqi Zhao	Fred Hutchinson Cancer Research Center
2021 IMS Program Chair	Lexin Li	University of California Berkeley

2021 WNAR Student Paper Competition Committee

Laura Saba (Chair)	University of Colorado Anschutz Medical Campus
Jarrett Barber	Arizona state University
Cindy Feng	Dalhousie University
Camille Moore	National Jewish Health
Holly Steeves	University of Victoria
Mourad Tighiouart	Cedars-Sinai Medical Center
Julie Zhou	University of Victoria

2021 WNAR Operations Committee

Audrey Hendricks (Chair)	University of Colorado Denver
Jennifer McNichol	University of New Brunswick
Lingling An	University of Arizona
Subodh Selukar	University of Washington

Presidential Invited Address

Rachel Fewster, PhD

Professor of Statistics
University of Auckland, New Zealand

Bio

Rachel Fewster started life as a nature-mad youngster. After studying maths at Cambridge, UK, she returned to nature with a PhD in statistical ecology at St Andrews. In 1999 she moved to New Zealand for a two-year postdoc, and is somehow still there. She works in all aspects of ecological statistics, from applied to mathematical, and runs the citizen science project CatchIT involving several thousand members of the NZ public. She is also an enthusiastic educator, and holds two national teaching awards. In her spare time she loves exploring the magnificent wilderness areas of New Zealand, including what may qualify as some of its most cobwebby corners, and connecting with its distinctively quirky wildlife.

Title: How to count the things you didn't see: the magic and mystery of estimating population size

For those who wish to save a species, wipe out a virus, eliminate crime, or engage in a host of other worthy pursuits, one question is often at the forefront: how many are there, of the things we wish to save or eliminate? Unfortunately, those things that we seek to count typically do not seek to be counted, so a measure of cunning is required. The art of estimating population size is to leverage whatever information we can glean from those items we did see, to tell us how many more items we didn't see.

I will outline traditional approaches to estimating population size, and show why a goldmine of interesting problems still remains for modern statistical researchers. Most of the examples will be drawn from wildlife population scenarios, in which new survey technologies ranging from satellite imagery to acoustic recordings leave us unsure even of how many animals we did detect, let alone how many we didn't. Overcoming the challenges posed by these big, enigmatic data streams will enable a massive upscaling of wildlife and biodiversity monitoring across some of the world's most inaccessible environments. However, new ways of thinking are called for, and I will describe some recent progress.

Short Course 1

Master Protocols: Tracking Complex Disease (and COVID-19!) with Bayesian Adaptive Platform Trials

Instructor: **Ben Saville,** Berry Consultants
Anna McGlothlin, Berry Consultants
Christina Saunders, Berry Consultants

Course Description:

As medical research continues to push into new frontiers of discovery and personalized patient care, along with new complex diseases and worldwide pandemics (COVID-19), it is imperative that clinical trial designs and statistical methodologies evolve to address the forthcoming challenges. One key innovation is the master protocol, including "platform" trial designs which can evaluate multiple therapies simultaneously in complex heterogenous diseases. In this course, we explain Bayesian adaptive methodologies underlying modern trials with master protocols. We introduce fundamental concepts in Bayesian adaptive trials, including Bayesian priors and posteriors, predictive probabilities, hierarchical modeling and "basket" trials, adaptive sample size, and response adaptive randomization. We explain the objectives and efficiencies of adaptive platform trial designs, with high profile examples investigating treatments in COVID-19, Amyotrophic Lateral Sclerosis (ALS), and Cancer. We show the role of virtual trial simulation in trial design, and discuss logistical and practical considerations in the implementation of these complex designs. In addition, we discuss the impact of the COVID-19 pandemic on both design and implementation of adaptive clinical trials. A highlight of the course will be interactive breakout activities that encourage individual participation and teach key adaptive platform trial concepts. Upon completion of the course, participants will have a general understanding of Bayesian adaptive platform trials and underlying methodologies, and better recognize opportunities for innovation in their respective organizations.

Short Course 2

Mediation Analysis and Software with Applications to Explore Health Disparities

Instructor: **Lu Tian,** Stanford University

Dr. Tian is Professor at the Department of Biomedical Data Science of Stanford University. Lu Tian received his Sc.D. in Biostatistics from Harvard University. He has considerable experience in statistical methodological research, planning large epidemiological studies, performing data management for randomized clinical trials and conducting applied data analysis. His current research interest includes developing statistical methods in survival analysis, semiparametric regression modelling, high-dimensional data analysis, precision medicine and meta-analysis. He has published more than 200 peer reviewed journal articles and currently served as the Associate Editor of Chance, Biometrics and Statistics in Medicine.

Course Description:

In a prospective clinical study to compare two groups, the primary end point is often the time to a specific event (for example, disease progression, death). The hazard ratio estimate is routinely used to empirically quantify the between-group difference under the assumption that the ratio of the two hazard functions is constant over time. When this assumption is plausible, such a ratio estimate may capture the relative difference between two survival curves. However, the clinical meaning of such a ratio estimate is difficult (if not impossible) to interpret when the underlying proportional hazards assumption is violated. In this course, we will discuss several critical concerns regarding this conventional practice and propose an attractive alternative for quantifying the underlying differences between groups based on restricted mean survival time (RMST). I will discuss various issues in employing RMST in practical analysis including statistical inference, result interpretation, selecting the truncation point, study design, power comparison, regression adjustment and extensions to competing risk and recurrent events settings. We will discuss the pros and cons of the RMST-based analysis and demonstrate that it is competitive to its hazard ratio-based conventional counterparts in many real world applications.

Scientific Program

Monday, June 14, 2021 8:30-10:15am PDT

Statistical learning and inference in online, dynamic settings

Organizer & Chair: Jean Feng, University of California, San Francisco

8:30 **Online non-parametric estimation and the quest for computational efficiency**
Noah Simon, University of Washington

8:55 **Online analysis of high-dimensional Gaussian graphical models**
George Michallidis, University of Florida

9:20 **Online Multiple Hypothesis Testing**
Tijana Zrnic, University of California, Berkeley

9:45 **Bayesian logistic regression for online recalibration and revision of risk prediction models with guarantees**
Jean Feng, University of California, San Francisco

Novel statistical approaches for disease early detection and diagnosis with complex data

Organizer & Chair: Ying Huang, Fred Hutchinson Cancer Research Center

8:30 **Challenges and Opportunities: Evaluation of Biomarkers for Early Risk Prediction, Early Detection and Diagnosis, and Prognosis**
Ziding Feng, Fred Hutchinson Cancer Research Center

8:55 **Estimation of Diagnostic Accuracy Based on Group-Tested Results**
Aiyi Liu, NICHD/NIH

9:20 **A multivariate parametric empirical Bayes screening approach for early detection of hepatocellular carcinoma using multiple longitudinal biomarkers**
Nabihah Tayob, Dana-Farber Cancer Institute

9:45 **Strategies for validating biomarkers using data from a reference set**
Ying Huang, Fred Hutchinson Cancer Research Center

Advancement of Roust Statistical Methods for Omics Studies

Organizer & Chair: Wen Zhou, Colorado State University

8:30 **AdaPT: An interactive procedure for multiple testing with side information**
Lihua Lei, Stanford University

8:55 **Large-scale inference of multivariate regression for heavy-tailed and asymmetric data**
Wen Zhou, Colorado State University

9:20 **Inference of Robust Regression with Contaminated Errors**
Zhao Ren, University of Pittsburgh

9:45 **Smoothed Quantile Regression**
Wenxin Zhou, University of California, San Diego

Recent developments in meta-analysis and data integration

Organizer & Chair: Lifeng Lin, Florida State University

8:30 **A Variance Shrinkage Method Improves Arm-Based Bayesian Network Meta-Analysis**
Haitao Chu, University of Minnesota

8:55 **A multivariate to multivariate approach for voxel-wise genome-wide association analysis**
Shuo Chen, University of Maryland School of Medicine

9:20 **A model for effect modification using targeted learning with observational data arising from multiple studies**
Mireille Schnitzer, University de Montreal

9:45 **Evaluation of various estimators for standardized mean difference in meta-analysis**
Lifeng Lin, Florida State University

Advances in ecological data modelling

Organizer & Chair: Hideyasu Shimadzu, Loughborough University, UK

8:30 **Integrating multiple sources of ecological data to estimate species abundance of woody plants at geographic scales**
Keiichi Fukaya, National Institute for Environmental Studies, Japan

8:55 **Estimating Abundance from Animal Traces**
Rafael Moral, Maynooth University, Ireland

9:20 **Categorical data analysis to investigate spatial and temporal trend for Integrated Ecosystem Assessment in the Norwegian Sea**
Hiroko Solvang, Institute of Marine Research, Norway

9:45 **Spatiotemporal modeling of an estuarine decapod using Bayesian inference: environmental drivers of juvenile blue crab abundance**
Grace Chiu, Virginia Institute of Marine Science

Monday, June 14, 2021 10:30-12:15pm PDT

Statistical Challenges in Analysis and Implementation of Results Using Electronic Health Records and Insurance Claims Data

Organizer & Chair: Menggang Yu, University of Wisconsin

10:30 **Handling Outcome Misclassification and Selection Bias in Association Studies Using Electronic Health Records**
Bhramar Mukherjee, University of Michigan

10:55 **Meeting the mandate of the 21st Century Cures Act: Overcoming the challenges of real world data to improve cancer care and outcomes**
Rebecca Hubbard, University of Pennsylvania

11:20 **Using medical insurance claims to measure structural features of both health organizations and the physicians within them to aid the study of variations in health care**
James O'Malley, Dartmouth College

11:45 **Improve patient identification for the University of Wisconsin health system's complex case management program**

Menggang Yu, University of Wisconsin

Biomarkers, Prediction, and Clinical Outcomes: Applications in Kidney Transplant and Disease

Organizer & Chair: Kathleen Kerr, University of Washington

10:30 **Quantifying Overall Donor Effects on Transplant Outcomes Using Kidney Pairs from Deceased Donors**
Kathleen Kerr, University of Washington

10:55 **Development and assessment of risk models for interval-censored events post kidney transplant using the variability of a longitudinal biomarker**
Kristen Campbell, University of Colorado

11:20 **Prediction of atrial fibrillation in chronic kidney disease**
Leila Zelnick, University of Washington

11:45 **Assessing the Impacts of Misclassified Case-Mix Factors on Healthcare Provider Profiling: performance of dialysis facilities**
Yi (Lisa) Mu, Actelion Pharmaceuticals

Bayesian methods for incorporating external data in clinical trials

Organizer: Ben Saville, Berry Consultants

Chair: Christina Saunders, Berry Consultants

10:30 **Bayesian methods for incorporating external control data in adaptive clinical trials**
Kristian Thorlund, McMaster University

10:55 **Bayesian borrowing of external treatment effect: A recent FDA device approval in heart failure**
Ben Saville, Berry Consultants

11:20 **Bayesian Sequential Monitoring for Pediatric Clinical Trials with Adult Data Extrapolation**
Matt Psioda, University of North Carolina at Chapel Hill

11:45 **Quantifying the amount of information added to a trial from the incorporation of external data**
Kert Viele, Berry Consultants

New Developments on Statistical Learning and Inference

Organizer & Chair: Linglong Kong, University of Alberta

10:30 **Proximal Temporal Consistent Learning for Estimating Infinite Horizon Dynamic Treatment Regimes**
Ruoqing Zhu, University of Illinois at Urbana-Champaign

10:55 **An integrated model approach to activation signatures and backgroung connectivity for task fMRI data**
Michelle Miranda, University of Victoria

11:20 **Adaptive-to-model hybrid test for regressions**
Lingzhu Li, University of Alberta

11:45 **Causal Inference Using Sufficient Dimension Reduction**
Yeying Zhu, University of Waterloo

Student paper presentation 1

Chair: Laura Saba, University of Colorado

10:30 **Profile Matching for the Generalization, Transportation, and Personalization of Causal Inferences**
Eric Cohn, Harvard University

10:50 **On the implied weights of linear regression for causal inference**
Ambarish Chattopadhyay, Harvard University

11:10 **Nonparametric causal mediation analysis for stochastic interventional (in)direct effects**
Nima Hejazi, University of California, Berkeley

11:30 **Accurate Risk Prediction for Cardiovascular Disease Intervention across Multiple Subpopulations from UK Biobank Data**
Waverly Wei, University of California, Berkeley

Monday, June 14, 2021 12:15-1:15pm PDT

Graduate Student Social Hour

Monday, June 14, 2021 1:45-3:30pm PDT

Innovative Statistical Methodology Development in Precision Medicine

Organizer & Chair: Lei Liu, Washington University in St. Louis

1:45 **Testing a high-dimensional parameter in the presence of high-dimensional nuisance parameters**
Wei Pan, University of Minnesota

2:10 **Causal inference via artificial neural networks: from prediction to causation**
Shujie Ma, UC Riverside

2:35 **New Approaches for Inference on Optimal Treatment Regimes**
Lan Wang, University of Miami

3:00 **Precision Medicine: Interaction survival tree approach for recurrent event data**
Lei Liu, Washington University in St. Louis

Statistical inference in modern, large-scale time series data

Organizer: Shizhe Chen, University of California, Davis

Chair: Xu Shi, University of Michigan

1:45 **An Instrumental Variable Method for Point Processes**
Shizhe Chen, University of California, Davis

2:10 **Time-varying overlapping clustering method via latent factor model**
Kean Ming Tan, University of Michigan

2:35 **Causal Inference on Distribution Functions**
Linbo Wang, University of Toronto

3:00 **On Proximal Causal Inference With Synthetic Controls**
Xu Shi, University of Michigan

The 200th Birth Anniversary of Florence Nightingale: Celebrating Women in Statistics - Past, Present, and Future.

Organizer & Chair: Nusrat Jahan, James Madison University

1:45 **Empowering women in statistics for 50 years: History of the Caucus for Women in Statistics**
Tomi Mori, St. Jude Children's Research Hospital

2:10 **Paving the Way: Women as Mentors and Advocates for Junior Statisticians**
Jessica Minnier, OHSU-PSU School of Public Health

2:35 **Statistics Education in the field of Health Sciences**
Nusrat Jahan, James Madison University

3:00 **Florence Nightingale Day: Inspiring and Passing the 'Lamp' to the Next Generation Statisticians**
Shili Lin, Ohio State University

Recent Advances in Neuroimaging Analysis

Organizer: Lexin Li, University of California, Berkeley
Chair: Jingshen Wang, University of California, Berkeley

1:45 **Brain connectivity-informed regularization methods in multi-modal imaging**
Jaroslaw Harezlak, Indiana University

2:10 **A Bayesian approach to joint modeling of matrix-valued imaging data and treatment outcome with applications to depression studies**
Bei Jiang, University of Alberta

2:35 **Time-varying l_0 optimization for Spike Inference from Multi-Trial Calcium Recordings**
Zhaoxia Yu, University of California, Irvine

3:00 **Functional Response Quantile Regression Model**
Linglong Kong, University of Alberta

New directions in radiation epidemiology

Organizer: Munechika Misumi, Radiation Effects Research Foundation
Chair: Richard Sposto, Radiation Effects Research Foundation

1:45 **Radiation risk estimation and statistical methods for the long-term follow-up studies of Japanese Atomic Bomb Survivors**
Munechika Misumi, Radiation Effects Research Foundation

2:10 **An application of multiple indicators, multiple causes measurement error models to adjust for dose error in RERF data**
Carmen Tekwe, Indiana University Bloomington

2:35 **Biologically based models of cancer risk in radiation research**
Jan Christian Kaiser, Helmholtz zentrum münchen Germany

3:00 **Statistical issues in estimating factors affecting the individual response to radiation**
Kyoji Furukawa, Kurume University

Monday, June 14, 2021 3:45-5:30pm PDT

Topics in Causal Inference

Organizer: Lexin Li, University of California, Berkeley
Chair: Kuang-Yao Lee, Temple University

3:45 **Inference for algorithm-agnostic variable importance**
Marco Carone, University of Washington

4:10 **Causal Estimation in Observational Data Subject to Missing by A Machine Learning Approach**
Xiaochun Li, Indiana University

4:35 **Inference on Heterogeneous Quantile Treatment Effects via Rank-Score Balancing**
Jingshen Wang, University of California, Berkeley

5:00 **Floor Discussion**

Design and modeling for complex featured data

Organizer: Zhezhen Jin, Columbia University
Chair: Lu Tian, Stanford University

3:45 **Design and analysis of biomarker-integrated clinical trials with adaptive threshold detection and flexible patient enrichment**
Ting Wang, Biogen

4:10 **Incorporating Imaging in Cure Rate Models**
Zhangsheng Yu, Shanghai Jiaotong University

4:35 **Semi-/non-parametric regression for pooled response data**
Xianzheng Huang, University of South Carolina

5:00 **Dynamic Risk Prediction Triggered by Intermediate Events Using Survival Tree Ensembles**
Yifei Sun, Columbia University

New fronts in survival and longitudinal data analysis in biomedical research

Organizer: Zhigang Li, University of Florida
Chair: Yimei Li, St. Jude Children's research hospital

3:45 **An efficient implementation of a semiparametric joint model for longitudinal and competing risks data**
Gang Li, UCLA

4:10 **Joint Penalized Spline Modeling of Multivariate Longitudinal Data**
Lihui Zhao, Northwestern University

4:35 **Sample Size Estimation for Trials of Recurrent Events with Additive Treatment Effects**
Liang Zhu, University of Texas at Houston

5:00 **Joint modeling in presence of informative censoring in palliative care studies**
Zhigang Li, University of Florida

Modern Methods in Ecological Statistics

Organizer & Chair: Laura Cowen, University of Victoria

3:45 **Multi-Year Bayesian Hierarchical Framework to Smoothly Fill Missing Data Gaps in Mark-Recapture Studies**
Audry Beliveau, University of Waterloo

4:10 **Mark-recapture and Bayesian State Space Analysis of Fish Movements in the Region of Canadian Arctic**
Saman Muthukumarana, University of Manitoba

4:35 **The role of computation in estimating abundance of large carnivores in Scandinavia**
Perry De Valpine, University of California - Berkeley

5:00 **Maximum unified fatty acid signature analysis: A novel approach to QFASA**
Holly Steeves, University of Western Ontario

Student paper presentation 2

Chair: Mourad Tighiouart, Cedars-Sinai

3:45 **Using decision theory to avoid fallacies of post hoc power**
Chloe Krakauer, University of Washington and Kaiser Permanente Washington Health Research Institute

4:05 **The relationship between the Bayes factor and separation of Bayesian credible intervals in within-subject designs**
Zhengxiao Wei, University of Victoria

4:15 **Bayesian Variance Estimation and Hypothesis Testing Using Inference Loss Functions**
Kendrick (Qijun) Li, University of Washington

4:35 **Quantifying uncertainty in spikes estimated from calcium imaging data**
Yiqun Chen, University of Washington

Tuesday, June 15, 2021 8:30-10:15am PDT

Novel statistical methods for Personalized Treatments

Organizer & Chair: Bibhas Chakraborty, National University of Singapore

8:30 **Assessing dynamic treatment regimes embedded in a SMART with an ordinal outcome**
Bibhas Chakraborty, National University of Singapore

8:55 **Estimation and inference on high-dimensional individualized treatment rule in observational data using split-and-pooled de-correlated score**
Yingqi Zhao, Fred Hutchinson Cancer Research Center

9:20 **Some comparisons between likelihood and surrogate based objective functions for individualized treatment rule estimation**
Michael Kosorok, University of North Carolina at Chapel Hill

9:45 **Discussant:** Eric Laber, Duke University

Analytical Methods for Time to Event Endpoints with Non-proportional Hazards

Organizer: Amarjot Kaur, Merck Research Labs
Chair: Qing Li, Merck Research Labs

8:30 **A Robust Design Approach for Clinical Trials with Potential Non-proportional Hazards: A Straw Man Proposal**
Satrajit Roychoudhury, Pfizer Inc.

8:55 **A user's perspective on the analytical methods under non-proportional hazards**
Amarjot Kaur, Merck Research Labs

9:20 **Weighted Kaplan-Meier statistics and RMST – the better testing and estimating strategy for time to event analysis in study with fixed duration?**
Zhiliang Li, CRISPR Therapeutics

9:45 **Efficiency vs. Interpretability in Clinical Trials Testing**
Richard Chappell, University of Wisconsin, Madison

Complex functional data analysis

Organizer: Kuang-Yao Lee, Temple University
Chair: Bei Jiang, University of Alberta

8:30 **Cross-Component Registration for Multivariate Functional Data, With Application to Growth Curves**
Hans Mueller, University of California at Davis

8:55 **Unified Principal Component Analysis for Sparse and Dense Functional Data under Spatial Dependency**
Yehua Li, University of California at Riverside

9:20 **Hypothesis testing for functional linear models**
Yu-Ru Su, Kaiser Permanente Washington Health Research Institute

9:45 **Functional sufficient dimension reduction through average Fréchet derivatives**
Kuang-Yao Lee, Temple University

Change-point detection, inference, and applications to biological data

Organizer: Ning Hao, University of Arizona
Chair: Yue Niu, University of Arizona

8:30 **A Nonparametric procedure for Frechet Change point detection**
Yichao Wu, University of Illinois at Chicago

8:55 **Change point localization in dependent dynamic nonparametric random dot product graphs**
Oscar Madrid Padilla, University of California at Los Angles

9:20 **Equivariant Variance Estimation for Multiple Change-point Model**
Han Xiao, Rutgers University

9:45 **A super scalable algorithm for short segment detection**
Yue Niu, University of Arizona

Recent Methods for Analyzing Infectious Disease Data

Organizer & Chair: Peihua Qiu, University of Florida

8:30 **A longitudinal Bayesian mixed effects model with hurdle Conway-Maxwell-Poisson distribution**

Jeremy Gaskins, University of Louisville

8:55 **Statistical Adjustment for Reporting Bias in Outbreak Data of Infectious Diseases**
Yang Yang, University of Florida

9:20 **Effective Spatio-Temporal Surveillance of Infectious Diseases**
Kai Yang, University of Florida

9:45 **Effective Spatio-Temporal Surveillance of Infectious Diseases**
Peihua Qiu, University of Florida

Tuesday, June 15, 2021 10:30-12:15pm PDT

Analysis of wearable devices data in biomedical studies

Organizer & Chair: Ken Wang, Fred Hutchinson Cancer Research Center

10:30 **Eliciting longitudinal physical activity patterns using densely sampled accelerometry**
Loki Natarajan, University of California, San Diego

10:55 **Streamlining the collection and pre-processing of accelerometry data in large cohort studies and clinical trials**
Jacek Urbanek, Johns Hopkins University

11:20 **Graph-based tests on mean and variance components of the repeatedly assessed physical activity density objects**
Haochang Shou, University of Pennsylvania

11:45 **Functional data analysis methods for characterizing physical activity intensity and duration using accelerometry data**
Chongzhi Di, Fred Hutchinson Cancer Research Center

Valid statistical approaches in non-randomized oncology study data analysis

Organizer & Chair: Sunhee Ro, Sierra Oncology

10:30 **Accounting for Patient Selection in the Interpretation of Single Arm Phase 2 trials**

Eric Holmgren, BeiGene Pharmaceuticals USA

10:55 **How to Make a "Relatively Fair" Comparison without a Randomized Controlled Trial**
Zhiyue Huang, Roche

11:20 **A randomized Phase II design which brings in information on the control arm from past studies to reduce the sample size**
Mithat Gonen, Memorial Sloan Kettering Cancer Center

11:45 **Propensity-score based vs regression based approach for adjusting bias in treatment effect estimate from non-randomized, cross-trial comparison**
Sunhee Ro, Sierra Oncology

Recent Statistical Developments in High-Dimensional Omics Sciences

Organizer & Chair: Debmalya Nandy, University of Colorado Anschutz Medical Campus

10:30 **Differential Expression Analysis using Kernel Machines for CyTOF data**
Tusharkanti Ghosh, University of Colorado Anschutz Medical Campus

10:55 **Identifying condition-specific patterns in large-scale genomic data**
Qunhua Li, Penn State University

11:20 **Towards Mechanistic Inferences in Radiomics**
Debashis Ghosh, University of Colorado Anschutz Medical Campus

11:45 **scClassify: multiscale classification of cells using single and multiple reference**
Jean Yee Hwa Yang, University of Sydney

Contributions to spatio-temporal models with applications to environmental and ecological data

Organizer & Chair: Claudio Fuentes, Oregon State University

10:30 **Multivariate spatial analysis of non-negative responses using SF-NNGPs**
Daniel Taylor-Rodriguez, Portland State University

10:55 **Spatial Modeling of Zero-Inflated Data with Copula Models**
Lisa Madsen, Oregon State University

11:20 **Nonparametric Spatio-Temporal Hawkes Processes: Benefits and Uses**
James Molyneux, Oregon State University

11:45 **A linear mixed model formulation for spatio-temporal random processes with computational advances for the product, sum, and product--sum covariance functions**
Michael Dumelle, Pacific Ecological Systems Division - EPA

Student paper presentation 3

Chair: Jarrett Barber, Northern Arizona University

10:30 **Summix: A method for detecting and adjusting for population structure in genetic summary data**
Ian Arriaga-MacKenzie, University of Colorado

10:50 **Efficiency and precision for hidden population models**
Matthew Parker, Simon Fraser University

11:10 **Doubly robust capture-recapture methods for estimating population size**James
Manjari Das, Carnegie Mellon University

11:30 **A Tree-based Federated Learning Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources.**
Xiaoqing Tan, University of Pittsburgh

11:50 **A Tree-based Federated Learning Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources.**
Xinyuan Dong, University of Washington

Tuesday, June 15, 2021 12:15-1:30pm PDT

Early Career Panel

Organizer: Ying Lu, Stanford University

Moderator: Ying Lu, Stanford University

12:15 **Discussion**

Panelists: Nebiyou Bekele, Excelisis

Brad Biggerstaff, CDC
Chito Hernandez, Biomarin
Joan Hu, Simon Fraise University
Karen Messer, University of California San Diego
Megan Othus, Fred Hutchinson Cancer Research Center

Tuesday, June 15, 2021 1:45-3:30pm PDT

Advances in statistical approaches for handling High-dimensional data

Organizer & Chair: MinJae Lee, University of Texas Southwestern

1:45 **ConQuR-ing Batch Effects in Microbiome Profiling Studies using Conditional Quantile Mapping**
Michael Wu, Fred Hutchinson Cancer Research Center

2:10 **Extensions of machine learning methods for classification of objects based on high-dimensional measurements of embedded observations within each object**
Jose-Miguel Yamal, University of Texas School of Public Health

2:35 **Integrative Analysis of Multi-Omic Data via Sparse Multiple Co-Inertia Analysis**
Qi Long, University of Pennsylvania

3:00 **A Landscape of Acquired Allelic Imbalance across the Cancer Continuum**
Paul Scheet, The University of Texas MD Anderson Cancer Center

Recent developments in functional data analysis

Organizer & Chair: Chongzhi Di, Fred Hutchinson Cancer Research Center

1:45 **Joint model for survival and multivariate sparse functional data with application to a study of Alzheimer's Disease**
Luo Xiao, North Carolina State University

2:10 **Modeling trajectories using functional linear first-order differential equations**
Julia Wrobel, University of Colorado Denver

2:35 **Robust functional principal components for sparse longitudinal data**
Matias Salibian-Barrera, University of British Columbia

3:00 **Robust Functional Principal Component Analysis viaA Functional Pairwise Spatial Sign Operator**
Ken Wang, Fred Hutchinson Cancer Research Center

Novel methods in latent class analysis

Organizer: Sarah Schmiege, University of Colorado Anschutz Medical Campus

Chair: Mary Sammel, University of Colorado Anschutz Medical Campus

1:45 **Multilevel Latent Class Analysis for Cross-Classified Data Structures**

Katherine Masyn, University of Illinois at Chicago

2:10 **Latent Class Analysis with Time-Varying Covariate Effects: A Simulation Study and Empirical Example of LCA-TVEM**
Bethany C. Bray, University of Colorado Denver

2:35 **Confirmatory latent class methods evaluating performance of threshold boundary and equality constraints**
Sarah Schmiege, University of Colorado Anschutz Medical Campus

3:00 **Joint latent class modeling approach for predicting clinical outcomes with longitudinal profiles of biomarkers subject to limits of detection**
Lan Kong, Penn State University College of Medicine

Student paper presentation 4

Chair: Holly Steeves, Western University

1:45 **Fitting a stochastic model of intensive care occupancy to noisy hospitalization time series**
Achal Awasthi, Duke University

2:05 **Effectiveness of Localized Lockdowns in the COVID-19 Pandemic**
Yige Li, Harvard T. H. Chan School of Public Health

2:25 **RECeUS: Ratio Estimation of Censored Uncured Subjects, A Different Approach for Studying Sufficient Follow-Up in Studies of Long-Term Survivors**
Subodh Selukar, University of Washington

2:45 **Ant Colony System Optimization for spatiotemporal Modelling of Combined EEG and MEG Data**
Eugene Opoku, University of Victoria

Contributed paper presentation 1

Chair: Yingqi Zhao, Fred Hutchinson Cancer Research Center

1:45 **(Poster Presentation) New selection method to identify pleiotropic variants associated with both quantitative and qualitative traits**
Kipoong Kim, Pusan National University

1:55 **(Poster Presentation) Effectiveness of Localized Lockdowns in the COVID-19 Pandemic**
Xianglong Liang, Pusan National University

2:05 **Simulating Bugs Over Time: A User-Friendly Guide to Simulating Longitudinal OTU Counts Using the Dirichlet-Multinomial Distribution**
Nicholas Weaver, University of Colorado

2:20 **Extension of the Condition-adaptive Fused Graphical Lasso and Application to Modeling Brain Region Co-Expression Networks**
Souvik Seal, University of Colorado

2:35 **The Impact of Continuity Corrections on Rare-Event Meta-Analysis**
Brinley Zabriskie, Brigham Young University

2:50 **Extension of the Two-Step Approach for Informative Dropout in Survival Analysis**
Cristina Murray-Krezan, University of New Mexico

3:05 **Salmon stock forecasting using remote sensing covariates**
Mehnaz Jahid, University of Victoria

Tuesday, June 15, 2021 3:45-5:00pm PDT

Presidential Invited Address

Organizer & Chair: Laura Cowen, University of Victoria

3:45 **How to count the things you didn't see: the magic and mystery of estimating population size**
Rachel Fewster, The University of Auckland

Wednesday, June 16, 2021 8:30-10:15am PDT

Frontiers of statistical genomics: deep learning and beyond

Organizer & Chair: Wei Sun, Fred Hutchinson Cancer Research Center

8:30 **Knockoff genotypes: value in counterfeit**
Chiara Sabatti, Stanford University

8:55 **Integrating GWAS and multi-omics QTL summary statistics to elucidate disease genetic mechanisms via a hierarchical low-rank model**
Lin Chen, University of Chicago

9:20 **A new clustering algorithm for assigning cells to known cell types according to marker genes**
Jun Li, University of Notre Dame

9:45 **DeepGWAS to Enhance GWAS Signals for Neuropsychiatric Disorders via Deep Neural Network**
Yun Li, UNC Chapel Hill

Statistical Considerations for N-of-1 Clinical Trial Designs

Organizer & Chair: Sonia Jain, University of California, San Diego

8:30 **nof1: an R package for analyzing and presenting n-of-1 trials**
Jiabei Yang, Brown University

8:55 **Modeling Individual Goal Achievement Behavior Using Bayesian Networks**
Christian Pascual, University of California, San Diego

9:20 **Statistical considerations of Bayesian Model Parameters Under Fixed or Random Intercepts**
Kexin Qu, Brown University

9:45 **A Bayesian-bandit adaptive design for N-of-1 clinical trials**
Sama Shrestha, Pharmapace, Inc

Analysis of health outcomes data with complex correlation structures

Organizer: Ann Lazar, University of California San Francisco

Chair: Elizabeth Juarez-Colunga, University of Colorado Denver

8:30 **Use of copulas for analyzing discrete longitudinal and clustered data**
Rao Chaganty, Old Dominican University

8:55 **Correlated gap time analysis with flexible hazards applied to pulmonary exacerbations in the EPIC Observational Study**

Elizabeth Juarez-Colunga, University of Colorado Denver

9:20 **The mixed model for repeated measures for cluster randomized trials**
Melanie Bell, The University of Arizona

9:45 **Discussant**: Ann Lazar, University of California San Francisco

Contributed paper presentation 2

Chair: Xinyuan Dong, University of Washington

8:30 **Improving Random Forest Predictions in Small Datasets from Two-phase Sampling Designs**
Sunwoo Han, Fred Hutchinson Cancer Research Center

8:45 **Random Forests for Time Series Forecasting and Forecast Intervals**
Barbara Bailey, San Diego State University

9:00 **SurvBenchmark: comprehensive benchmarking study of survival analysis methods using both omics data and clinical data**
Yunwei Zhang, The University of Sydney

9:15 **A High-dimensional Mediation Model for a Neuroimaging Mediator: Integrating Clinical, Neuroimaging, and Neurocognitive Data to Mitigate Late Effects in Pediatric Cancer**
Jade Xiaoqing Wang, St. Jude Children's Research Hospital

9:30 **Accurate Source Tracking Using Microbial Samples with Applications in Forensic Study**
Qianwen Luo, The University of Arizona

9:45 **Calibration Coefficient Estimation in Quantitative Fatty Acid Signature Analysis**
Jennifer McNichol, The University of New Brunswick

Student paper presentation 5

Chair: Cindy Feng, Dalhousie University

10:30 **Development of an augmented high-dimensional graphical lasso model to incorporate prior biological knowledge for global network learning**
Yonghua Zhuang, University of Colorado Anschutz Medical Campus

10:50 **Learning network formation patterns with stochastic block models**
Zitong Zhang, University of California Davis

11:10 **Precision Matrix Estimation under the Horseshoe-like Prior-Penalty Dual**
Sagar Ksheera, Purdue University

11:30 **High-dimensional semi-supervised learning: in search of optimal inference of the mean**
Yuqian Zhang, University of California, San Diego

Wednesday, June 16, 2021 10:30-12:15pm PDT

High-dimensional inference with applications to -omics data

Organizer & Chair: Tusharkanti Ghosh, University of Colorado, Anschutz Medical Campus

10:30 **CCmed: Cross-condition mediation analysis for identifying replicable trans-associations mediated by cis-gene expression**
Fan Yang, University of Colorado, Anschutz Medical Campus

10:55 **An Exploration of Multiple-Testing Correction Methods in Large-Scale Omics Studies**
Debmalya Nandy, University of Colorado Anschutz Medical Campus

11:20 **Mechanism-Aware Imputation: A two-step approach in handling missing values in metabolomics**
Elin Shaddox, University of Colorado, Anschutz Medical Campus

11:45 **Compositional Data Analysis using Kernels in Mass Cytometry Data**
Pratyaydipta Rudra, Oklahoma State University

Recent Advancements in Spatio-Temporal Modeling

Organizer & Chair: Ali Arab, Georgetown University

10:30 **Multivariate spatio-temporal models for landscape change using aerial imagery**
Xinyi (Lucy) Lu, Colorado State University

10:55 **Conjugate spatio-temporal Bayesian multinomial Polya-gamma regression for the reconstruction of climate using pollen**
John Tipton, University of Arkansas

11:20 **A Bayesian approach for estimating age-adjusted rates for low-prevalence diseases over space and time**
Melissa Jay, University of Iowa

11:45 **Strategies for Modeling Dynamics of Emerging Epidemics**
Ali Arab, Georgetown University

Recent advances in designs and quantitative analysis in immunological research

Organizer & Chair: Tao He, San Francisco State University

10:30 **Characterization of the landscape of repertoire sequencing data with novel statistical approaches and advanced machine learning techniques**
Li Zhang, University of California, San Francisco

10:55 **Alternative Analysis Methods for Non-proportional Hazards in Cancer Immunology Studies**
Ray Lin, Genetech

11:20 **Design for immuno-oncology clinical trials involving non-proportional hazards patterns**
Zhenzhen Xu, FDA

11:45 **Floor Discussion**

Recent Development in Interrupted Time Series Methods

Organizer: Maricela Cruz, Kaiser Permanente Washington Health Research Institute

Chair: Michelle Nuno, University of South California

10:30 **Power and sample size calculation for interrupted time series analyses of count outcomes**
Shangyuan Ye, Harvard Pilgrim Health Care Institute and Harvard Medical School

10:55 **A formal test for the existence of a change point in Interrupted Time Series**
Maricela Cruz, Kaiser Permanente Washington Health Research Institute

11:20 **An Improved Method for Analysis of Interrupted Time Series (ITS) Data: Accounting for Patient and Healthcare Heterogeneity Using Weighted Analysis**
Joycelyne Ewusle, University of Ottawa

11:45 **Birds of a feather flock together: Comparing controlled pre-post designs** Ali Arab, Carrie Fry, Vanderbilt University

Student paper presentation 6

Chair: Camile Moore, National Jewish Health

10:30 **Scale mixture of skew-normal linear mixed models with within-subject serial dependence**
Fernanda Lang Schuumacher, University of Campinas

10:50 **REHE: Fast Variance Components Estimation for Linear Mixed Models**
Kun Yue, University of Washington

11:10 **Exact inference for fixed-effects meta-analysis of proportions**
Spencer Hansen, University of Washington

11:30 **Power Analysis for Stepped Wedge Trials with Multiple Interventions**
Phillip Sundin, University of California Los Angeles

11:50 **Model misspecification in stepped wedge trials: Random effects for time or treatment**
Emily Voldal, University of Washington

Abstracts

Invited Sessions Sponsored by WNAR

Novel statistical approaches for disease early detection and diagnosis with complex data

Organizer: Ying Huang, Biostatistics, Bioinformatics & Epidemiology Program, Fred Hutchinson Cancer Research Center

Strategies for validating biomarkers using data from a reference set

Ying Huang (Fred Hutchinson Cancer Research Center), Lu Wang (Peking University), Ziding Feng (Fred Hutchinson Cancer Research Center)*

Candidate biomarkers discovered in the laboratory need to be rigorously validated before advancing to clinical application. However, it is often expensive and time-consuming to collect the high quality specimens needed for validation; moreover, such specimens are often limited in volume. The Early Detection Research Network has developed valuable specimen reference sets that can be used by multiple labs for biomarker validation. To optimize the chance of successful validation, it is critical to efficiently utilize the limited specimens in these reference sets on promising candidate biomarkers. Towards this end, we propose a novel two-stage validation strategy that partitions the samples in the reference set into two groups for sequential validation and rotates group membership to maximize the usage of available samples. We develop analytical formulas for performance parameters of this strategy in terms of the expected numbers of biomarkers that can be evaluated and the truly useful biomarkers that can be successfully validated, which can provide valuable guidance for future study design.

Estimation of Diagnostic Accuracy Based on Group-Tested Results

Wei Zhang (Chinese Academy of Sciences), Aiyi Liu (National Institute of Child Health Health and Human Development), Qizhai Li (Chinese Academy of Sciences), Paul Albert (National Cancer Institute)*

This talk concerns the problem of estimating a continuous distribution in a diseased or nondiseased population when only group-based test results on the disease status are available. The problem is challenging in that individual disease statuses are not observed and testing results are often subject to misclassification, with further complication that the misclassification may be differential as the group size and the number of the diseased individuals in the group vary. We propose a method to construct nonparametric estimation of the distribution and obtain its asymptotic properties. The performance of the distribution estimator is evaluated under various design considerations concerning group sizes and classification errors.

A multivariate parametric empirical Bayes screening approach for early detection of hepatocellular carcinoma using multiple longitudinal biomarkers

Nabihah Tayob (Dana-Farber Cancer Institute), Anna Lok (University of Michigan), Ziding Feng (Fred Hutchinson Cancer Research Center)*

The early detection of hepatocellular carcinoma (HCC) is critical to improving outcomes since advanced HCC has limited treatment options. Blood-based biomarkers are a promising direction since they are more easily standardized and less resource intensive than standard of care imaging. Combining of multiple biomarkers is more likely to achieve the sensitivity required for a clinically useful screening algorithm and the longitudinal trajectory of biomarkers contains valuable information that should be utilized. We propose a multivariate parametric empirical Bayes (mPEB) screening approach that defines personalized thresholds for each patient at each screening visit to identify significant deviations that trigger additional testing with more sensitive imaging. The

Hepatitis C Antiviral Long-term Treatment against Cirrhosis (HALT-C) trial provides a valuable source of data to study HCC screening algorithms. We study the performance of the mPEB algorithm applied to serum alpha-fetoprotein, a widely used HCC surveillance biomarker, and des-gamma carboxy prothrombin, an HCC risk biomarker that is FDA approved but not used in practice in the United States.

Advancement of Roust Statistical Methods for Omics Studies

Organizer: Wen Zhou, Colorado State University

Smoothed Quantile Regression

Xuming He (University of Michigan), Xiaoou Pan (UCSD), Kean Ming Tan (University of Michigan), Wenxin Zhou (UCSD)*

Quantile regression is a powerful tool for learning the relationship between a response variable and a multivariate predictor while exploring heterogeneous effects. In this talk, we consider statistical inference for quantile regression with large-scale data in the “increasing dimension” regime. We provide a systematic analysis of a convolution-type smoothing approach that achieves adequate approximation to computation and inference for quantile regression. This method, which we refer to as conquer, admits fast and scalable gradient-based algorithms to perform optimization, and multiplier bootstrap for statistical inference. Extensions to high-dimensional settings will also be discussed. Software implementing the methodology is available in the R package conquer.

Statistical Inference of Robust Regression with Contaminated Errors

Zhao Ren (University of Pittsburgh), Peiliang Zhang (University of Pittsburgh), Wen-Xin Zhou (University of California San Diego)*

We study the robust estimation and inference problems for linear regression in the increasing dimension regime. Given random design, we consider the conditional distributions of error terms are contaminated by some arbitrary distribution (possibly depending on the covariates) with proportion \epsilon but otherwise can also be heavy-tailed and asymmetric. We show that simple robust M-estimators such as Huber and smoothed Huber, with an additional intercept added in the model, can achieve the minimax rates of convergence under the l2 loss. In addition, two types of confidence intervals with root-n consistency are provided by a multiplier bootstrap technique when the necessary condition on contamination proportion \epsilon=o(1/\sqrt{n}) holds. For a larger \epsilon, we further propose a debiasing procedure to reduce the potential bias caused by contamination, and prove the validity of the debiased confidence interval. At last, we extend our methods to the communication-efficient distributed estimation and inference setting. A comprehensive simulation study exhibits the effectiveness of our proposed inference procedures.

Large-scale inference of multivariate regression for heavy-tailed and asymmetric data

Youngseok Song (Colorado State University), Wen Zhou (Colorado State University), Wen-Xin Zhou (University of California San Diego)*

We consider the fundamental statistical problem of simultaneously testing a large number of general linear hypotheses, encompassing covariate-effect analysis, analysis of variance, and model comparisons. The new challenge that comes along with the overwhelmingly large number of tests is the ubiquitous presence of heavy-tailed and/or highly skewed measurement noise, which is the main reason for the failure of conventional least squares based methods. For large-scale multivariate regression, we develop a set of robust inference methods to explore data features, such as heavy tailedness and skewness, which are invisible to the scope of least squares. The new testing procedure is built on data-adaptive Huber regression, and a new covariance estimator of the regression estimate. Under mild conditions, we show that the proposed methods produce

consistent estimates of the false discovery proportion. Extensive numerical experiments, along with an empirical study on quantitative linguistics, demonstrate the advantage of our proposal compared to many state-of-the-art methods when the data are generated from heavy-tailed and/or skewed distributions.

AdaPT: An interactive procedure for multiple testing with side information
Lihua Lei (Stanford University), Will Fithian (University of California, Berkeley)*

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

Change-point detection, inference, and applications to biological data
Organizer: Ning Hao, University of Arizona

A Nonparametric procedure for Frechet Change point detection
Miaomai Zhou (University of Illinois at Chicago), Yichao Wu (University of Illinois at Chicago)*

We consider the change point detection for data situated in a generic metric space, which is called Frechet change point detection. A nonparametric procedure is proposed to perform Frechet change point detection. We provide theoretical support for the proposed method and demonstrate its competitive finite-sample performance with both simulated data and real data.

Change point localization in dependent dynamic nonparametric random dot product graphs
Oscar Hernan Madrid Padilla (University of California, Los Angeles)*

We study the change point localization problem in a sequence of dependent nonparametric random dot product graphs. To be specific, assume that at every time point, a network is generated from a nonparametric random dot product graph model. The underlying distributions are piecewise constant in time and change at unknown locations, called change points. Most importantly, we allow for dependence among networks generated between two consecutive change points. This setting incorporates edge-dependence within networks and temporal dependence between networks, which is the most flexible setting in the published literature. To accomplish the task of consistently localizing change points, we propose a novel change point detection algorithm, consisting of two steps. First, we estimate the latent positions of the random dot product model. Subsequently, we construct a nonparametric version of the CUSUM statistic (Page, 1954, Padilla et al., 2019) that allows for temporal dependence. Consistent localization is proved theoretically and supported by extensive numerical experiments, which illustrate state-of-the-art performance

Equivariant Variance Estimation for Multiple Change-point Model
Ning Hao (The University of Arizona), Yue Niu (The Unviersity of Arizona), Han Xiao (Rutgers University)*

The variance of noise plays an important role in many change-point detection procedures and the associated inferences. Most commonly used variance estimators require strong assumptions on the true mean structure or normality of the error distribution, which may not hold in applications. More importantly, the qualities of these estimators have not been discussed systematically in the literature. In this paper, we introduce a framework of equivariant variance estimation for multiple change-point models. In particular, we characterize the set of all equivariant unbiased quadratic variance estimators for a family of change-point model classes, and develop a minimax theory for such estimators.

Recent developments in meta-analysis and data integration

Organizer: Lifeng Lin, Department of Statistics, Florida State University

A Variance Shrinkage Method Improves Arm-Based Bayesian Network Meta-Analysis

Zhenxun Wang (University of Minnesota), Lifeng Lin (Florida State University), James S Hodges (University of Minnesota), Richard Maclehose (University of Minnesota), Haitao Chu (University of Minnesota)*

Network meta-analysis (NMA) is a commonly used tool to combine direct and indirect evidence in systematic reviews of multiple treatments. Unlike the contrast-based NMA approach, which focuses on estimating relative effects such as odds ratios, the arm-based (AB) NMA approach can estimate absolute effects, which are arguably more informative. However, the number of clinical studies involving each treatment is often small, leading to unstable treatment-specific variance estimates in the AB-NMA approach when using non- or weakly-informative priors under an unequal variance assumption. Additional assumptions, such as homogeneous variances for all treatments, may be used to remedy this problem but such assumptions may be inappropriately strong. This article introduces a variance shrinkage method for an AB-NMA. Specifically, we assume different treatment variances share a common prior with unknown hyper-parameters. This assumption improves estimation by shrinking the variances in a data-dependent way. We illustrate the advantages of the variance shrinkage method by re-analyzing an NMA of organized inpatient care interventions for stroke and comprehensive simulations.

Evaluation of various estimators for standardized mean difference in meta-analysis

Lifeng Lin (Florida State University), Ariel Aloe (University of Iowa)*

Meta-analyses (MAs) of a treatment's effect compared with a control frequently calculate the meta-effect from standardized mean differences (SMDs). SMDs are usually estimated by Cohen's d or Hedges' g. Cohen's d divides the difference between sample means of a continuous response by the pooled standard deviation, but is subject to nonnegligible bias for small sample sizes. Hedges' g removes this bias with a correction factor. The current literature is confusingly inconsistent about methods for synthesizing SMDs. Using conventional methods, the variance estimate of SMD is associated with the point estimate of SMD, so Hedges' g is not guaranteed to be unbiased in MAs. This talk reviews and evaluates available methods for synthesizing SMDs. Their performance is compared using simulation studies and analyses of actual datasets. Because of the intrinsic association between point estimates and standard errors, the usual version of Hedges' g can result in more biased meta-estimation than Cohen's d. We recommend using average-adjusted variance estimators to obtain an unbiased meta-estimate, and the Hartung-Knapp-Sidik-Jonkman method for accurate estimation of its confidence interval.

A model for effect modication using targeted learning with observational data arising from multiple studies

Yan Liu (McGill University), Mireille Schnitzer (Université de Montréal), Guanbo Wang (McGill University), Edward Kennedy (Carnegie Mellon University), Dick Menzies (McGill University Health Centre), Andrea Benedetti (McGill University)*

When the effect of treatment may vary by individual, precision medicine can be improved by identifying patient covariates to predict the effect at the individual level. One may impose a working model in order to smooth (or summarize) the conditional effect rather than estimate the effect separately for all possible patient subgroups. When working with observational data one must also adjust for all potential confounders of the treatment-outcome relationship, which can be accomplished with propensity score and/or outcome regression modeling. Due to large data requirements, investigators may be interested in using the individual patient data from multiple studies. Our data arise from a systematic review of observational studies contrasting different treatment regimens for patients with multidrug-resistant tuberculosis, where multiple antibiotics are taken concurrently over a long period to cure the infection. We develop a doubly robust targeted learning (TMLE) method to fit a marginal structural model representing the treatment effect model in the individual patient data network meta-analytic setting when, for instance, any given treatment may not be observed in all studies.

A multivariate to multivariate approach for voxel-wise genome-wide association analysis
Qiong Wu (University of Maryland), Tianzhou Ma (University of Maryland), Peter Kochunov (University of Maryland), Yuan Zhang (Ohio State University), Shuo Chen (University of Maryland)*

The integrative analysis of imaging-genetics data facilitates the systematic investigation of genetic effects on brain structures and functions with spatial specificity. We focus on voxel-wise genome-wide association analysis, which may involve trillions of SNP-voxel pairs. We attempt to identify underlying organized association patterns of SNP-voxel pairs and understand the polygenic and pleiotropic networks on brain imaging traits. We first show that the probability of a non-trivial bi-clique (i.e., a set of SNPs highly correlated with a cluster of voxels) converges to zero under the null. Next, we develop computational strategies to detect latent SNP-voxel bi-cliques and inference model for statistical testing. We validate our method by extensive simulation studies, and then apply it to a voxel-wise genome-wide association analysis based on genetic data and white matter integrity data of 110 participants from the human connectome project (HCP). We identify systematic effects of genetic variants on brain voxels in SNP-voxel bi-cliques on multiple chromosomes.

Innovative Statistical Methodology Development in Precision Medicine

Organizer: Lei Liu, Washington University in St. Louis

Causal inference via artificial neural networks: from prediction to causation
Xiaohong Chen (Yale University), Ying Liu (University of California, Riverside), Shujie Ma (University of California-Riverside), Zheng Zhang (Renmin University of China)*

Recent technological advances have created numerous large-scale datasets in observational studies, which provide unprecedented opportunities for evaluating the effectiveness of various treatments. Meanwhile, the complex nature of large-scale observational data post great challenges to the existing conventional methods for causality analysis. In this talk, I will introduce a new unified approach that we have proposed for efficiently estimating and inferring causal effects using artificial neural networks. We develop a generalized optimization estimation through moment constraints with the nuisance functions approximated by artificial neural networks. This general optimization framework includes the average, quantile and asymmetric least squares treatment effects as special cases. The proposed methods take full advantage of the large sample size of large-scale data and provide effective protection against mis-specification bias while achieving dimensionality

reduction. We also show that the resulting treatment effect estimators are supported by reliable statistical properties that are important for conducting causal inference.

New Approaches for Inference on Optimal Treatment Regimes
Lan Wang (University of Miami)*

Finding the optimal treatment regime (or a series of sequential treatment regimes) based on individual characteristics has important applications in precision Medicine. We propose two new approaches to quantify uncertainty in optimal treatment regime estimation. First, we consider inference in the model-free setting, which does not require to specify an outcome regression model. Existing model-free estimators for optimal treatment regimes are usually not suitable for the purpose of inference, because they either have nonstandard asymptotic distributions or do not necessarily guarantee consistent estimation of the parameter indexing the Bayes rule due to the use of surrogate loss. We study a smoothed robust estimator that directly targets the parameter corresponding to the Bayes decision rule for optimal treatment regimes estimation. Next, we consider the high-dimensional setting and propose a semiparametric model assisted approach for simultaneous inference. Simulations results and real data examples are used for illustration.

Testing a high-dimensional parameter in the presence of high-dimensional nuisance parameters
Wei Pan (University of Minnesota)*

A key to personalized medicine is to detect interactions between a treatment and individual-specific characteristics, the latter of which is represented by a high-dimensional vector of demographic, clinical and genetic information (with induced high-dimensional nuisance parameters). Although there are some machine learning algorithms for this purpose, often they cannot be used for statistical inference, e.g. statistical significance testing, on such detected interactions. We develop such a test and show its performance and application.

Precision Medicine: Interaction survival tree approach for recurrent event data
Chamila Perera (Washington University in St. Louis), Xiaogang Su (University of Texas, El Paso), Lei Liu (Washington University in St. Louis)*

Individual subjects can experience recurrent events in many biomedical studies. In this paper, we propose an interaction survival tree (IT) approach to identify subgroups with heterogeneous treatment effects in comparative studies that involve recurrent event data. That is, some of the resulting subgroups may receive beneficial treatment effects while others may receive negligible or even negative effects. The proposed method follows the standard CART methodology by recursively partitioning the data into subsets that show the greatest interaction with the treatment. The heterogeneity of treatment effects is assessed through the Cox's proportional hazards model with gamma frailty. The resultant final tree structure is used to explore the overall interaction between treatment and other covariates. Both simulated experiments and an analysis of the CGD data set from a placebo-controlled randomized trial of gamma interferon on chronic granulomatous disease are provided for evaluation and illustration of the proposed procedure.

Biomarkers, Prediction, and Clinical Outcomes: Applications in Kidney Transplant and Disease

Organizer: Kathleen Kerr, University of Washington

Development and assessment of risk models for interval-censored events post kidney transplant using the variability of a longitudinal biomarker
Kristen Campbell (University of Colorado Anschutz Medical Campus), Elizabeth Juarez-Colunga (University of Colorado Anschutz Medical Campus)*

This talk discusses methods for using the variability of a longitudinal biomarker to dynamically predict an interval-censored time to event outcome. We first investigate a shared random effects model with longitudinal and interval censored survival sub-models. In our motivating clinical example, the biomarker values were highly variable, and the higher the variance meant the patient was likely being non-adherent to treatment. Thus, individual variance of the longitudinal biomarker was thought to be important in prediction of adverse events. The shared random effects model incorporates the sharing of an individual-specific variance component, along with a traditional intercept and slope. Using this model, we develop a dynamic prediction framework to calculate individualized predicted probabilities of event-free survival for new subjects, based on historical biomarker measurements and demographic data.

Prediction of atrial fibrillation in chronic kidney disease

Leila Zelnick (University of Washington), Michael Shlipak (University of California San Francisco), Elsayed Z. Soliman (Wake Forest University), Amanda Anderson (Tulane University), Robert Christenson (University of Maryland), James Lash (University of Illinois-Chicago), Rajat Deo (University of Pennsylvania), Panduranga Rao (University of Michigan), Farsad Afshinnia (University of Michigan), Jing Chen (Tulane University), Jiang He, (Tulane University), LA Stephen Seliger (University of Maryland), Ray Townsend (University of Pennsylvania), Debbie L. Cohen (University of Pennsylvania), Alan Go (Kaiser Permanente Northern California), Nisha Bansal (University of Washington)*

Atrial fibrillation (AF) is common in chronic kidney disease (CKD) and associated with poor outcomes; prediction models may identify high risk patients. We compared a published AF prediction model with new machine learning (ML) models in a CKD population. We studied 2766 AF-free participants in the CRIC cohort study with complete cardiac biomarker and clinical data, evaluating the utility of several ML methods and a previously validated model (CHARGE-AF) to predict AF. Discrimination was assessed using Harrell's C-index; calibration was also evaluated. Mean (SD) age was 57 (11) years, 55% men, 38% black, and mean (SD) eGFR 45 (15) mL/min/1.73m2; 259 incident AF events occurred during follow-up. The CHARGE-AF equation performed similarly to a boosting model with clinical data only; adding NT-proBNP significantly improved the C-index. In addition to NT-proBNP and hsTnT, the final model included age, black race, Hispanic ethnicity, CVD, COPD, MI, PVD, use of ACEi/ARBs, calcium channel blockers, diuretics, height, and weight. Using ML algorithms, a model which included 12 clinical variables and 2 cardiac biomarkers had moderate discrimination of incident AF in a CKD population.

Quantifying Overall Donor Effects on Transplant Outcomes Using Kidney Pairs from Deceased Donors

Kathleen F. Kerr (University of Washington), Eric R. Morenz (University of Washington), Heather Thiessen Philbrook (Johns Hopkins Medicine), F. Perry Wilson (Yale School of Medicine), Peter P. Reese (University of Pennsylvania Perelman School of Medicine), Chirag R. Parikh (Johns Hopkins Medicine)*

In kidney transplantation, it is unknown how much donor factors contribute to good or bad outcomes versus other factors such as recipient characteristics and protocol differences. Kidney transplants from deceased donors offer an opportunity to address this question, since a deceased donor provides two kidneys that are transplanted into different recipients. The opportunity presented by kidney pairs has been previously recognized but not fully exploited. We adapted methods used in genetics, twin studies, and other fields with paired data to quantify the impact of donor effects on kidney transplant outcomes. We propose new concordance metrics that are more relevant to the clinical context of kidney transplant. Overall results indicate that donor factors have small or moderate impact on post-transplant clinical outcomes.

Assessing the Impacts of Misclassified Case-Mix Factors on Healthcare Provider Profiling: performance of dialysis facilities

Yi Mu (Nektar Therapeutics), Andrew Chin (University of California, Davis), Abhijit Kshirsagar (University of North Carolina, Chapel Hill), Heejung Bang (University of California, Davis)*

Quantitative metrics are used to develop profiles of healthcare institutions, including hospitals, nursing homes and dialysis clinics. However, there is some concern about how misclassification in case-mix factors, which are typically accounted for in profiling, impacts results. We evaluated the potential effect of misclassification on profiling results, using 20,744 patients from 2,740 dialysis facilities in the US Renal Data System. We compared 30-day readmission as the profiling outcome measure, using comorbidity data from either the CMS Medical Evidence Report (error-prone) or Medicare claims (more accurate). Although the regression coefficient of the error-prone covariate demonstrated notable bias in simulation, the outcome measure (standardized readmission ratio) and profiling results were quite robust. Thus, we conclude that misclassification on case-mix did not meaningfully impact overall profiling results. We also identified both extreme degree of case-mix factor misclassification and magnitude of between-provider variability as two factors that can potentially exert enough influence on profile status to move a clinic from one performance category to another.

Bayesian methods for incorporating external data in clinical trials

Organizer: Ben Saville, Berry Consultants

Bayesian Sequential Monitoring for Pediatric Clinical Trials with Adult Data Extrapolation

Mathew Psioda (UNC Chapel Hill), Xiaoqiang Xue (Dova Pharmaceuticals)

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 are often available before pediatric trials begin. Motivated by these common aspects of pediatric drug development, 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 comparatively 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 as well as principles for rational decision making that are critical when information in the pediatric data is not substantial, making prior-data conflict difficult to discern.

Bayesian borrowing of external treatment effect: A recent FDA device approval in heart failure

Ben Saville (Berry Consultants)*

Heart failure occurs when a person's heart is unable to pump sufficient blood to meet the body's needs, and is one of the leading causes of hospitalizations and deaths in the United States. The FDA recently approved an implantable device for restoring a normal timing pattern of the heartbeat based on evidence from a randomized clinical trial demonstrating patient benefit. The primary analysis used a Bayesian repeated measures model for the change in peak oxygen update, in which "Bayesian borrowing" formally incorporated an estimated treatment effect from an external randomized clinical trial of the same experimental device. This presentation will include details on trial design, the Bayesian Power Prior methodology for borrowing, statistical and clinical justification, regulatory considerations, and modeling results of the trial data. This includes discussion of the FDA "breakthrough devices" program, as well as interactions with the FDA panel involved in the regulatory approval.

Statistical Challenges in Analysis and Implementation of Results Using Electronic Health Records and Insurance Claims Data

Organizer: Menggang Yu, University of Wisconsin

Meeting the mandate of the 21st Century Cures Act: Overcoming the challenges of real world data to improve cancer care and outcomes

Rebecca Hubbard (University of Pennsylvania), Kylie Getz (University of Pennsylvania), Ronac Mamtani (University of Pennsylvania)*

Randomized clinical trials (RCTs) are considered the gold standard for evaluating treatment efficacy but suffer from limitations. Vulnerable populations are under-represented in RCTs, raising concerns about equity and the external validity of results. Clinical trial treatment settings may not reflect care and outcomes as they are experienced in routine practice. Real-world data (RWD) sources have the potential to address many of these concerns. However, limitations of RWD necessitate careful attention to study design and application of appropriate statistical methods. In this talk, I will discuss the potential of RWD to supplement RCT evidence on treatment efficacy. I will present novel statistical developments aimed at extending evidence on treatment efficacy beyond clinical trials by combining clinical trial and EHR-derived effectiveness estimates. Through judicious application of these methods RWD have the potential to aid clinical decision making through the creation of evidence on treatment effectiveness that would not be possible using RCTs alone.

Using medical insurance claims to measure structural features of both health organizations and the physicians within them to aid the study of variations in health care

James O'malley (Dartmouth)*

In this presentation I will describe a general procedure for forming large-scale physician social networks using only administrative medical insurance claims data. Methods for summarizing the structure of the network in terms of a diverse range of social network and network science metrics will be outlined for both the entire network and meaningful subnetworks, such as those based on hospital affiliation or region. A second set of metrics are used to represent each physician's position within any network in which they are an actor (or node). Finally, multi-level statistical models are specified and used to regress important patient-level or hospital-level health-related dependent variables on the various types of network measures and any other measured predictors to determine whether the network variables account for previously unexplained variation in the health variables. Care is taken to interpret the fitted models, especially if the network is developed on the same or a related data set.

Improve patient identification for the University of Wisconsin health system's complex case management program

Menggang Yu (University of Wisconsin), Jared Huling (University of Minnesota)*

There has been a great interest in developing interventions to effectively coordinate the typically fragmented care of patients with many comorbidities. Since 2013, the University of Wisconsin health system implemented a complex case management intervention for such purpose. We found out that the intervention effect was quite differential among the diverse patient population. Given the resource constraint, it is imperative to identify which patients may benefit the most from the intervention. We accomplish such goal by modeling effect-modifying covariates from electronic health records and insurance claims. Instead of using the traditional outcome-modeling approach, we propose a general framework for scoring intervention benefits. In particular, we construct personalized scores ranking the patients according to their potential treatment effects.

Recent Methods for Analysing Infectious Disease Data

Organizer: Peihua Qiu, University of Florida

A longitudinal Bayesian mixed effects model with hurdle Conway-Maxwell-Poisson distribution

Tong Kang (University of Florida), Jeremy Gaskins (University of Louisville), Steven Levy (University of Iowa), Somnath Datta (University of Florida)*

Dental caries (i.e., cavities) is one of the most common chronic childhood diseases and may progress throughout a person's lifetime. The Iowa Fluoride Study (IFS) was designed to investigate the effects of various fluoride, dietary and nondietary factors on the progression of dental caries among a cohort of Iowa children. We develop a mixed effects model to perform a analysis of the longitudinal clustered data of IFS at ages 5, 9, 13, and 17. We combine a Bayesian hurdle framework with the Conway-Maxwell-Poisson regression model, which can account for both excessive zeros and various levels of dispersion. A hierarchical shrinkage prior distribution is used to share the temporal information for predictors in the fixed-effects model. The dependence among teeth of each child is modeled through a sparse covariance structure of the random effects across time. Simulation studies are conducted to assess the accuracy and effectiveness of our statistical methodology. The results of this article provide novel tools to statistical practitioners and offer fresh insights to dental researchers on effects of various risk and protective factors on caries progression.

Effective Disease Surveillance By Using Covariate Information

Kai Yang (University of Florida)*

Effective surveillance of infectious diseases is critically important for public health and safety of our society. Incidence data of such diseases are often collected spatially from multiple clinics and hospitals through some disease reporting systems. In such a system, new batches of data keep being collected over time, and a decision needs to be made immediately after new data are collected regarding whether there is a disease outbreak at the current time point. This is the disease surveillance problem that will be focused in this talk. There are some existing methods for solving this problem, most of which use the disease incidence data only. In practice, however, disease incidence is often associated with some covariates. I will introduce a new methodology which can make use of the helpful covariate information to improve its effectiveness. A novelty of our method is behind its property that only the covariate information associated with a true disease outbreak can help trigger a signal from the proposed method. Our method can accommodate seasonality, spatio-temporal data correlation and nonparametric data distribution, making it feasible to use in many real applications.

Effective Spatio-Temporal Surveillance of Infectious Diseases

Kai Yang (University of Florida), Peihua Qiu (University of Florida)*

Effective surveillance of infectious diseases is critically important for public health. Governments have spent much resource in building national and regional disease reporting and surveillance systems. In these systems, conventional control charts, such as the cumulative sum (CUSUM) and the exponentially weighted moving average (EWMA) charts, are usually included for disease surveillance purposes. However, these charts require many assumptions on the observed data, including the ones that the observed data are independent at different locations and/or times, and they follow a parametric distribution when no disease outbreaks are present. These assumptions are rarely valid in practice, making the results from the conventional control charts unreliable. We present a new sequential monitoring approach in this talk, which can accommodate the dynamic nature of the observed disease incidence rates (i.e., the distribution of the observed

disease incidence rates can change over time due to seasonality and other reasons), spatio-temporal data correlation, and nonparametric data distribution.

Statistical Adjustment for Reporting Bias in Outbreak Data of Infectious Diseases
Yang Yang (University of Florida), Mingjin Liu (University of Florida), Neda Jalali (University of Florida)*

Reporting bias is common in the surveillance of infectious diseases and takes different forms. A typical example is that outbreaks are investigated only if the number of cases exceeds a prespecified threshold. Another example is that most outbreak investigations only survey cases, ignoring individuals who are at risk but not infected, to which we refer as a missing denominator problem. Such bias, if left unaddressed, can lead to erroneous estimation of key epidemiological parameters. To adjust for the selection bias associated with reporting threshold, the likelihood for the transmission dynamic need to be conditioned on the observation that final size of the outbreak exceeds the threshold. However, traditional final size algorithms soon become numerically unstable even for moderate sizes. For the missing denominator problem, branching processes are often used, but the assumption of independent offspring distributions is questionable. We will discuss methods for addressing these challenges, simulation studies for validating these methods, and their application to real surveillance data of influenza and MERS-CoV.

New Developments on Statistical Learning and Inference

Organizer: Linglong Kong, University of Alberta

Causal Inference Using Sufficient Dimension Reduction
Wei Luo (Zhejiang University), Yeying Zhu (University of Waterloo), Debashis Ghosh (Colorado School of Public Health)*

In this talk, the use of sufficient dimension reduction (SDR) approaches for causal inference will be discussed. Compared with the original covariates and the propensity scores, which are commonly used in the causal inference literature, the reduced covariates obtained from SDR are non-parametrically estimable and are effective in imputing the missing potential outcomes, under a mild assumption on the low-dimensional structure of the data. The corresponding IPW, matching and double-robust estimators will be provided. The consistency of the proposed approaches require a weaker common support condition than the traditional approaches. We develop relevant asymptotic results and conduct simulation studies as well as real data analysis to illustrate the usefulness of the proposed approaches.

Adaptive-to-model hybrid test for regressions
Lingzhu Li (University of Alberta), Xuehu Zhu (Xi'an Jiaotong University), Lixing Zhu (Hong Kong Baptist University)*

In model checking for regressions, nonparametric estimation-based tests usually have tractable limiting null distributions and are sensitive to oscillating alternative models, but suffer from the curse of dimensionality. In contrast, empirical processbased tests can, at the fastest possible rate, detect local alternatives distinct from the null model, yet are less sensitive to oscillating alternatives and rely on Monte Carlo approximation for critical value determination, which is costly in computation. We propose an adaptive-to-model hybrid of moment and conditional moment-based tests to fully inherit the merits of these two types of tests and avoid the shortcomings. Further, such a hybrid makes nonparametric estimation-based tests, under the alternatives, also share the merits of existing empirical process-based tests. The methodology can be readily applied to other kinds of data and construction of other hybrids. As a by-product in sufficient dimension reduction field, a study on residual related central mean subspace and central subspace for model adaptation is devoted to showing when alternative models can be indicated and when cannot.

Statistical inference in modern, large-scale time series data

Organizer: Shizhe Chen, University of California, Davis

An Instrumental Variable Method for Point Processes

Shizhe Chen (University of California, Davis)*

We propose an instrumental variable method for causal inference with point process treatment and outcome. We define causal quantities of interest and establish nonparametric identification results with a binary instrumental variable. We extend the traditional Wald estimation for point process treatment and outcome, and show that it should be performed after a Fourier transform and thus takes the form of deconvolution. We term this as the generalized Wald estimation and propose an estimation strategy based on well-established deconvolution methods. The proposed estimation strategy is applicable under many commonly-used models without requiring distributional assumptions on the unmeasured confounders. For empirical illustration, we conduct simulations and apply the proposed methodology to analyze the data from an experiment on mouse neuron activities.

Causal Inference on Distribution Functions

Zhenhua Lin (National University of Singapore), Dehan Kong (University of Toronto), Linbo Wang (University of Toronto)*

Understanding causal relationships is one of the most important goals of modern science. So far, the causal inference literature has focused almost exclusively on outcomes coming from the Euclidean space $\mathbb{R}^p$. However, it is increasingly common that complex datasets collected through electronic sources, such as wearable devices, cannot be represented as data points from $\mathbb{R}^p$. In this paper, we present a formal definition of causal effects for outcomes from the Wasserstein space of cumulative distribution functions, which in contrast to the Euclidean space, is non-linear. We develop doubly robust estimators and associated asymptotic theory for these causal effects. As an illustration, we use our framework to quantify the causal effect of marriage on physical activity patterns using wearable device data collected through the National Health and Nutrition Examination Survey.

On Proximal Causal Inference With Synthetic Controls

Xu Shi (University of Michigan)*

We consider evaluating the impact of an intervention when time series data on a single treated unit and multiple untreated units are observed, in pre- and post- treatment periods. The synthetic control method relaxes the parallel trend assumption on which difference-in-differences methods typically rely upon. The term “synthetic control” (SC) refers to a weighted average of control units that is built to match the treated unit’s pre-treatment outcome trajectory, such that the SC’s post-treatment outcome predicts the treated unit’s unobserved potential outcome under no treatment. Common practice to estimate the weights is to regress the pre-treatment outcomes of the treated unit on that of the control units using ordinary or weighted least squares. However, it has been shown that these estimators can be inconsistent. In this talk, we introduce a proximal causal inference framework for the synthetic control approach, and formalize identification and inference for the average treatment effect on the treated unit.

Time-varying overlapping clustering method via latent factor model

Yanxin Jin (University of Michigan), Yang Ning (Cornell University), Kean Ming Tan (University of Michigan)*

Clustering is an important tool in interdisciplinary research such as genomics and neuroscience. One ubiquitous assumption for most clustering methods is that each variable belongs only to one cluster, and such an assumption may be unrealistic in many scientific settings. In this talk, we will introduce a clustering procedure using latent factor model that allows overlapping clusters, i.e., each variable can belong to multiple clusters. In particular, we focus on developing a method for clustering variables on time-varying data with clusters changing across time. Our proposed method is also able to match the cluster labels across time. Theoretical guarantees are established consistent estimation of the clusters.

Contributions to spatio-temporal models with applications to environmental and ecological data

Organizer: Claudio Fuentes, Oregon State University

A linear mixed model formulation for spatio-temporal random processes with computational advances for the product, sum, and product--sum covariance functions

Michael Dumelle (United States Environmental Protection Agency), Jay M. Ver Hoef (United States National Oceanic and Atmospheric Administration), Claudio Fuentes (Oregon State University), Alix Gitelman (Oregon State University)*

To properly characterize a spatio-temporal random process, it is necessary to understand the process' dependence structure. It is common to describe this dependence using a single random error having a complicated covariance. Instead of using the single random error approach, we describe spatio-temporal random processes using linear mixed models having several random errors; each random error describes a specific quality of the covariance. This linear mixed model formulation is general, intuitive, and contains many commonly used covariance functions as special cases. We focus on using the linear mixed model formulation to express three covariance functions: product (separable), sum (linear), and product--sum. We discuss benefits and drawbacks of each covariance function and propose novel algorithms to efficiently invert their covariance matrices, even when every spatial location is not observed at every time point. Via a simulation study, we assess model performance and computational efficiency of these covariance functions when estimated using restricted maximum likelihood (likelihood-based) and Cressie's weighted least squares (semivariogram-based).

Nonparametric Spatio-Temporal Hawkes Processes: Benefits and Uses

James Molyneux (Oregon State University - Department of Statistics)*

Self-exciting point process provide a method for modeling the occurrence of spatio-temporal phenomenon in which the occurrence of an event at one point causes a temporary excitation of similar events to occur nearby in time or space-time. In seismology, self-exciting point processes often take on a parametric form based on the well-understood properties of aftershocks. In other fields, where a parametric form is infeasible, nonparametric Hawkes processes provide a flexible method for estimating the self-exciting triggering densities as well as declustering events into background and excited events.

Spatial Modeling of Zero-Inflated Data with Copula Models

Lisa Madsen (Oregon State University), Vicente Monleon (US Forest Service)*

The traditional geostatistical model uses properties of the multivariate normal distribution for modeling and prediction. If the spatial process is non-Gaussian, we can mimic the geostatistical approach by employing a Gaussian copula model. We present an application based on zero-inflated forest inventory data and explore methods of model estimation and spatial prediction.

Multivariate spatial analysis of non-negative responses using SF-NNGPs
Daniel Taylor Rodriguez (Portland State University), Andrew Finley (Michigan State University)*

Non-negative high-dimensional response vectors are pervasive in population ecology and environmental monitoring. Popular continuous metrics for species abundance, such as biomass, are of this form. When considering communities of species jointly, modeling is further complicated by the large number of species needed to represent communities across large spatial domains. With these considerations in mind we formulate a multivariate Gamma regression model based on an approximation to the latent Spatial Factor Nearest Gaussian Process (SF-NNGP). This approximation reduces the computational burden of NNGP's by clustering according to similarities in the distance matrices between neighbor sets. We discuss preliminary results showing the computational advantages of the method.

Design and modeling for complex featured data

Organizer: Zhezhen Jin, Columbia University

Semi-/non-parametric regression for pooled response data
Xianzheng Huang (University of South Carolina), Dewei Wang (University of South Carolina), Xichen Mou (University of Memphis)*

We propose local polynomial and partially linear estimators for the conditional mean of a continuous response when only pooled response data are collected under different pooling designs. Asymptotic properties of these estimators are investigated and compared. Extensive simulation studies are carried out to compare finite sample performance of the proposed estimators under various model settings and pooling strategies. We apply the proposed regression methods to two real-life applications to illustrate practical implementation and performance of the estimators for the mean function.

Dynamic Risk Prediction Triggered by Intermediate Events Using Survival Tree Ensembles
Yifei Sun (Columbia University), Sy Han Chiou (University of Texas at Dallas), Colin Wu (National Institutes of Health), Meghan Mcgarry (University of California San Francisco), Chiung-Yu Huang (University of California San Francisco)*

With the availability of massive amounts of data from electronic health records and registry databases, incorporating time-varying patient information to improve risk prediction has attracted great attention. To exploit the growing amount of predictor information over time, we develop a unified framework for landmark prediction using survival tree ensembles, where an updated prediction can be performed when new information becomes available. Compared to conventional landmark prediction with fixed landmark times, our methods allow the landmark times to be subject-specific and triggered by an intermediate clinical event. In our framework, both the longitudinal predictors and the event time outcome are subject to right censoring, and thus existing tree-based approaches cannot be directly applied. To tackle the analytical challenges, we propose a risk-set-based ensemble procedure by averaging martingale estimating equations from individual trees. The methods are applied to the Cystic Fibrosis Patient Registry data to perform dynamic prediction of lung disease in cystic fibrosis patients and to identify important prognosis factors.

Design and analysis of biomarker-integrated clinical trials with adaptive threshold detection and flexible patient enrichment
Ting Wang (Biogen), Xiaofei Wang (Duke University), Stephen L George (Duke University), Haibo Zhou (University of North Carolina at Chapel Hill)*

We propose a new adaptive threshold detection and enrichment design in which the biomarker threshold is adaptively estimated and updated by optimizing a trade-off between the size of the biomarker positive population and the magnitude of the treatment effect in that population. Enrichment is based on an enrollment criterion that accounts for the uncertainty in estimation of the threshold. Early termination for futility is allowed based on predictive success probability. Valid testing and estimation techniques for the treatment effect overall and inpatient subgroups are studied. Simulations and an example demonstrate advantages of the proposed design over existing designs.

New fronts in survival and longitudinal data analysis in biomedical research

Organizer: Zhigang Li, University of Florida

Sample Size Estimation for Trials of Recurrent Events with Additive Treatment Effects

Liang Zhu (Eisai Inc)*

This study discusses the design of clinical trials where the primary endpoint is a recurrent event with the focus on the sample size calculation. For the problem, a few methods have been proposed but most of them assume a multiplicative treatment effect on the rate or mean number of recurrent events. In practice, sometimes the additive treatment effect may be preferred or more appealing because of its intuitive clinical meaning and straightforward interpretation compared to a multiplicative relationship. In this paper, new methods are presented and investigated for the sample size calculation based on the additive rates model for superiority, non-inferiority, and equivalence trials. They allow for flexible baseline rate function, staggered entry, random dropout, and overdispersion in event numbers, and simulation studies show that the proposed methods perform well in a variety of settings. We also illustrate how to use the proposed methods to design a clinical trial based on real data.

An efficient implementation of a semiparametric joint model for longitudinal and competing risks data

Shanpeng Li (UCLA), Ning Li (UCLA), Jin Zhou (University of Arizona), Hua Zhou (UCLA), Gang Li (UCLA)*

Semiparametric joint models of longitudinal and survival data are computationally costly and their current implementations often do not scale well to large data. This paper investigates and addresses some computational barriers that are common in semiparametric shared random effects joint models of longitudinal and survival data. Specifically, we show that key factors leading to computational bottlenecks in a typical EM algorithm for such joint models include numerical integration, risk set calculation, standard error estimation, and choice of the initial values. We further illustrate that these bottlenecks can be effectively addressed with efficient algorithms to generate drastic speedups, reducing the run-time from days to minutes for large data. We illustrate the computational gains in comparison with existing algorithms and R packages on both simulated and real world data.

Joint modeling in presence of informative censoring in palliative care studies

Quran Wu (University of Florida), Zhigang Li (University of Florida)*

Joint modeling of longitudinal data such as quality of life data and survival data is important for palliative care researchers to draw efficient inference because joint modeling can account for the associations between those two types of data that are commonly seen in palliative care studies. Modeling quality of life on a retrospective time scale from death time makes it convenient for investigators to interpret the analysis results of palliative care studies with relatively short life expectancy. However, censoring of death times, especially informative censoring such as

informative dropouts, poses challenges for modeling quality of life on a retrospective time scale. We develop a novel joint modeling approach that can address the challenge by allowing informative censoring events to be dependent on patients' quality of life through a random effect. There are three submodels in our approach: a linear mixed effect model for the longitudinal quality of life, a frailty model for the death time and another frailty model for the informative censoring time.

Modern Methods in Ecological Statistics

Organizer: Laura Cowen, University of Victoria

Multi-Year Bayesian Hierarchical Framework to Smoothly Fill Missing Data Gaps in Mark-Recapture Studies

Audrey Beliveau (University of Waterloo), Will Atlas (Wild Salmon Center), Thomas Buehrens (Washington Department of Fish and Wildlife)*

We consider mark-recapture studies of migrating animal populations. As animals pass a specific location along their trajectory, daily counts are recorded, subject to imperfect detection. A subsample is marked and released where they came from. Marked individuals re-observed on subsequent days inform detection probability, allowing to expand the imperfect counts to estimate the true population size. A common application of this sampling strategy is migrating salmon populations. But when river flow is too high, the daily counts are too dangerous to perform which can result in several consecutive days of missing data. To address this issue, among others, we develop a Bayesian hierarchical framework that shares information across years via Bayesian penalized splines. In particular, the shape of the yearly migration curves is modeled as a Generalized Additive Model (GAM) of day, year and their interaction. This allows information to flow across years while accommodating between-year variations. The method is demonstrated on a dataset of Sockeye salmon collected at Koeye River (Canada) from 2014 to 2019.

Mark-recapture and Bayesian State Space Analysis of Fish Movements in the Region of Canadian Arctic

Saman Muthukumarana (University of Manitoba)*

The region of Canadian arctic is a cold and arid region and despite the harsh climate, the Arctic animal population and their movements are very diverse. Over diverse species of fish live in Canadian Arctic waters, they have mixed movements in fresh and salt water over the time. We use mark-recapture methods and Bayesian state space framework to study the fish movement and survival changes in different sites in the study areas. The estimation of survival probabilities for different regions of the study area using multi-state mark recapture models will also be discussed.

The role of computation in estimating abundance of large carnivores in Scandinavia

Perry De Valpine (University of California - Berkeley), Daniel Turek (Williams College), Cyril Milleret (Norwegian University of Life Sciences), Pierre Dupont (Norwegian University of Life Sciences), Richard Bischof (Norwegian University of Life Sciences)*

This talk will summarize computational methods using the NIMBLE package that enabled Bayesian estimation of spatial capture-recapture models for brown bears, gray wolves, and wolverines at the scale of Norway and Sweden (Bischof et al. 2020: DOI: 10.1073/pnas.2011383117). The data come from noninvasive genetic sampling of over 35,000 scats from over 6,000 individuals. MCMC efficiency was dramatically improved using NIMBLE's configurable and extensible MCMC system (Milleret et al. 2018: DOI: 10.1002/ece3.4751. Turek et al. 2021: DOI: 10.1002/ecs2.3385). Methods included localizing detection probability calculations, vectorizing distance computations, block sampling of activity center coordinates, avoiding unnecessary computations for data-augmented individuals not currently in the model, and re-writing

the detection likelihood to use a sparse data format. A problem that started with 30-days of computation ending in a crash using JAGS was eventually feasible in 5 minutes using customized NIMBLE code. Many of these advances are available in the nimbleSCR R package.

Analytical Methods for Time to Event Endpoints with Non-proportional Hazards

Organizer: Amarjot Kaur, Merck Research Labs

A user's perspective on the analytical methods under non-proportional hazards

Amarjot Kaur (Merck Research Labs), Qing Li (Merck Research Labs)*

There has been a wide interest and research conducted in recent years in identifying robust analytical method(s) when the proportional hazards assumption is violated for the time to event data in clinical trials. It is well known that the standard models such as Log-rank test and Cox proportional hazards model lose efficiencies when the proportional hazard assumption is violated. The extent of loss in efficiencies depends upon various factors including the type of non-proportionality, i.e., change in the treatment effect magnitude only (quantitative) or change also in directionality (qualitative) over time. The Log-rank test and Cox proportional hazards model are robust under minor violations but not for major violations in proportional hazards assumption. For major violations, the standard methods not only lose power in testing the treatment effect but also present ambiguity in interpreting the results. In this talk, we will discuss user's perspective in analysis and interpretation of the time-to-event data when the proportional hazards assumption is violated and discuss some practical considerations.

Weighted Kaplan-Meier statistics and RMST – the better testing and estimating strategy for time to event analysis in study with fixed duration?

Ziliang Li (CRISPR Therapeutics)*

Recent methodology advance in hypothesis test and effect estimation for time-to-event (TTE) analysis under non-proportional hazards (NPH) largely focus on the setting where TTE endpoint is the primary endpoint with study duration varied to accumulate the pre-determined number of events. In this setting, more emphasis is placed on applying highly sensitive methods to identify a signal, i.e., a difference in hazard rates, followed by a companion hazard rate(s) estimate. In a lesser discussed setting where TTE is a (key) secondary endpoint where study duration is fixed, RMST with study duration as the truncation time appears to be a natural alternative to quantify treatment effect (under NPH). However, prior simulation evaluations suggested testing based on RMST alone lacks power compared to other KM estimate-based testing method(s) such as the Weight KM (WKM) statistics. In this presentation, we will do a deeper dive to the WKM statistics and illustrate the WKM statistics + RMST combo could be the better testing/estimating strategy for TTE analysis in study with fixed duration (under NPH).

A Robust Design Approach for Clinical Trials with Potential Non-proportional Hazards: A Straw Man Proposal

Satrajit Roychoudhury (Pfizer Inc), Keaven M Anderson (Merck and Co, Inc), Jiabu Ye (Merck and Co, Inc), Pralay Mukhopadhyay (Otsuka America Pharmaceuticals, Inc.)*

Targeting the immune system to cure cancer has emerged as a promising treatment option for patients in recent years. However, this novel treatment poses new challenges in the study design and statistical analysis of clinical trials. A major challenge is the delayed onset of treatment effects due to the mechanism of immunotherapy which violates the proportional hazard (PH) assumption. It is often referred as the non-proportional hazard (NPH) problem. In contrast to the PH assumption, NPH constitutes a broad class of alternative hypotheses. The conventional log-rank test may suffer a significant power loss in NPH scenarios. This presentation will focus on an

alternative design and analysis approach for immune-oncology trials. The proposed approach is based on a combination of multiple Fleming-Harrington WLR tests and is referred as the MaxCombo test. The main objective the new approach is to provide robust power for primary analysis under different NPH scenarios. Real-life examples and simulation studies will be presented for illustration.

Efficiency vs. Interpretability in Clinical Trials Testing
Richard Chappell (University of Wisconsin), Mitchell Paukner (University of Wisconsin)*

This talk will address the choice of estimand in non-inferiority and superiority trials with particular emphasis on the tension between efficiency and interpretability. Noninferiority (or equivalence) trials are medical experiments on humans which attempt to show that one intervention is not too much inferior to another on some quantitative scale. Naturally a lot of attention is given to choice of the margin of inferiority, but much less to its scale. Since null hypotheses in superiority studies generally imply no effect, they are often identical or at least compatible when formulated on different scales. However, nonzero Deltas on one scale usually conflict with those on another. For example, the four hypotheses of arithmetic or multiplicative differences of either survival or hazard in general all mean different things for noninferiority studies. This can lead to problems in interpretation when the clinically natural scale is not a statistically convenient one. In addition to this, I will also weigh in on recent discussions concerning efficient and interpretable comparisons of survival endpoints in clinical trials.

Complex functional data analysis
Organizer: Kuang-Yao Lee, Temple University

Cross-Component Registration for Multivariate Functional Data, With Application to Growth Curves
Hans-Georg Müller (UC Davis), Cody Carroll (UC Davis), Alois Kneip (University of Bonn)*

Multivariate functional data are becoming ubiquitous with advances in modern technology. We propose and study a novel model for multivariate functional data where the component processes are subject to mutual time warping in the form of systematic phase variation across their time domains, implemented as shift-warping across the components of the multivariate functional data. Estimates for these shifts are identifiable, enjoy parametric rates of convergence and often have intuitive physical interpretations, all in contrast to traditional curve registration methods. We illustrate this methodology with data from the Zürich Longitudinal Growth Study.

Functional sufficient dimension reduction through average Fréchet derivatives
Kuang-Yao Lee (Temple University), Lexin Li (University of California at Berkeley)*

In this work, we propose a new method for function-on-function sufficient dimension reduction (SDR), where both the response and the predictor are a function. We first develop the notions of functional central mean subspace and functional central subspace, which form the population targets of our functional SDR. We then introduce an average Fréchet derivative estimator, which extends the gradient of the regression function to the operator level and enables us to develop estimators for our functional dimension reduction spaces. We show the resulting functional SDR estimators are unbiased and exhaustive, and more importantly, without imposing any distributional assumptions such as the linearity or the constant variance conditions that are commonly imposed by all existing functional SDR methods. We establish the uniform convergence of the estimators for the functional dimension reduction spaces, while allowing both the number of Karhunen-Loève expansions and the intrinsic dimension to diverge with the sample size. We demonstrate the efficacy of the proposed methods through both simulations and two real data examples.

Hypothesis testing for functional linear models
Yu-Ru Su (Kaiser Permanente Washington Health Research Institute), Chongzhi Di (Fred Hutchinson Cancer Research Center), Li Hsu (Fred Hutchinson Cancer Research Center)*

Functional data arise frequently in biomedical studies, where the association between a functional predictor and a scalar response variable is the primary interest. We focused on hypothesis testing for the functional association under the widely used framework of functional linear models (FLM). A popular approach to testing the functional effects is through dimension reduction by functional principal component (FPC) analysis. We investigated the power performance of the Wald-type test with varying thresholds in selecting the number of PCs for the functional covariates, and showed that the power is sensitive to the chosen thresholds. We then proposed a new method of ordering and selecting PCs to construct test statistics. The proposed method accounted for both the association with the response and the variation along each eigenfunction. We showed that the proposed test is more robust against the choice of threshold while being as powerful as, and often more powerful than, the existing method. We applied the proposed method to cerebral white matter tracts data obtained from a diffusion tensor imaging tractography study.

Novel statistical methods for Personalized Treatments
Organizer: Bibhas Chakraborty, National University of Singapore

Assessing dynamic treatment regimes embedded in a SMART with an ordinal outcome
Palash Ghosh (Indian Institute of Technology - Guwahati), Xiaoxi Yan (National University of Singapore), Bibhas Chakraborty (National University of Singapore)*

Sequential multiple assignment randomized trials (SMART) are used to construct and assess data-driven dynamic treatment regimes (DTRs) based on an individual's treatment and covariate history. In the extant literature, the majority of the analysis methodologies for SMART data assume a continuous, binary or time-to-event primary outcome. However, ordinal outcomes are also quite common in clinical practice, as we will illustrate through our motivating carbohydrate periodization SMART. In this talk, we will introduce the notion of generalized odds-ratio (GOR) to compare two dynamic treatment regimes embedded in a SMART with an ordinal outcome. We will propose a likelihood-based approach to estimate GOR from SMART data, and discuss the asymptotic properties of the estimate. Next, we will use GOR to derive a sample size formula, and validate it through simulation experiments. We will illustrate the methodology by analyzing data from the motivating carbohydrate periodization SMART. A freely available R Shiny App software will be presented to facilitate wide dissemination.

Some comparisons between likelihood and surrogate based objective functions for individualized treatment rule estimation
Michael Kosorok (University of North Carolina at Chapel Hill)*

In this presentation, we present an asymptotic analysis comparing maximum likelihood versus surrogate based objective function estimation of dynamic treatment regimes. For simplicity, we focus on the linear boundary setting. We establish consistency, rate of convergence, and the limiting distribution for both approaches over a range of sharpness of the boundary between those who should receive one treatment versus the other. Surprisingly, neither approach dominates across all degrees of sharpness.

Estimation and inference on high-dimensional individualized treatment rule in observational data using split-and-pooled de-correlated score

Muxuan Liang (Fred Hutchinson Cancer Research Center), Young-Geun Choi (Sookmyung Women's University), Yang Ning (Cornell University), Maureen Smith (University of Wisconsin-Madison), Yingqi Zhao (Fred Hutchinson Cancer Research Center)*

With the increasing adoption of electronic health records, there is an increasing interest in developing individualized treatment rules (ITRs), which recommend treatments according to patients' characteristics, from large observational data. However, there is a lack of valid inference procedures for ITRs developed from this type of data in the presence of high dimensional covariates. In this work, we develop a penalized doubly robust method to estimate the optimal ITRs from high dimensional data. A split-and-pooled de-correlated score is proposed to construct hypothesis tests and confidence intervals. Our proposal utilizes the data splitting to conquer the slow convergence rate of nuisance parameter estimations, such as non-parametric methods for outcome regression or propensity models. We establish the limiting distributions of the split-and-pooled de-correlated score test and the corresponding one-step estimator in high dimensional setting. Simulation and real data analysis are conducted to demonstrate the superiority of the proposed method.

Recent Advancements in Spatio-Temporal Modeling

Organizer: Ali Arab, Georgetown University

Multivariate spatio-temporal models for landscape change using aerial imagery

Xinyi (Lucy) Lu (Department of Statistics, Colorado State University), Mevin Hooten (Department of Statistics, Colorado State University)*

The Alaskan landscape has undergone substantial changes in recent decades, most notably the expansion of shrubs and trees across the Arctic. We developed a dynamic statistical model to quantify the impact of climate change on the structural transformation of ecosystems using remotely sensed imagery. Our model accommodates changes in temperature and precipitation to infer and predict rates of land cover transitions while accounting for spatio-temporal heterogeneity. Transition types are highly correlated at both plot and subplot levels in our study system, therefore we explicitly characterized multi-scale spatial correlation using Gaussian processes. Because imagery pairs were collected at irregular time intervals, we also modeled dynamic state probabilities that evolve annually using a hierarchical framework. We developed a Polya-Gamma representation of our model to improve computation. Our model facilitates inference on the response of ecosystem state probabilities to shifts in climate and can be used to project future land cover transitions under various climate scenarios.

A Bayesian approach for estimating age-adjusted rates for low-prevalence diseases over space and time

Melissa Jay (University of Iowa), Jacob Oleson (University of Iowa), Mary Charlton (University of Iowa), Ali Arab (Georgetown University)*

Age-adjusted rates are frequently used by epidemiologists to compare disease incidence and mortality across populations. In small geographic regions, age-adjusted rates computed directly from the data are subject to considerable variability and are generally unreliable. Therefore, we desire an approach for estimating age-adjusted rates that accounts for the excessive number of zero counts in disease mapping datasets, which are naturally present for low-prevalence diseases and are further innated when stratifying the dataset by age group. Bayesian modeling approaches are naturally suited to employ spatial and temporal smoothing to produce more stable estimates of age-adjusted rates for small areas. We propose a Bayesian hierarchical spatio-temporal hurdle model for counts and demonstrate how age-adjusted rates can be estimated from the hurdle model. We illustrate our modeling approach with an application to cancer mortality at the county

level and present a simulation study to evaluate the proposed approach on datasets with varying characteristics.

Strategies for Modeling Dynamics of Emerging Epidemics
Ali Arab (Georgetown University), Naresh Neupane (Georgetown University), Ari Goldbloom-Helzner (ari_goldbloom-helzner@alumni.brown.edu)*

The dynamics of emerging epidemics are complex and difficult to understand and thus, challenging to model. Moreover, data for rare conditions (over time and space) often include excess zeros which may result in inefficient inference and ineffective prediction for such processes. This is a common issue in modeling rare or emerging diseases as well as early onset of infectious diseases or diseases that are not common in specific areas, specific time periods, or those conditions that are hard to detect. A common approach to modeling data with excess zeroes is to use zero-modified models (i.e., hurdle and zero-inflated models).Here, we discuss strategies for effectively model the dynamics of disease spread based on zero-modified hierarchical modeling approaches. To demonstrate our work, we provide a case study of modeling the spread of Lyme disease based on confirmed cases of the disease in the United States.

Conjugate spatio-temporal Bayesian multinomial Polya-gamma regression for the reconstruction of climate using pollen
John Tipton (University of Arkansas)*

One of the most-widely available climate proxy data are tree pollen collected in sediments. Pollen grains in sediments are counted and the relative abundance of different tree species is a function of the underlying climate state. Thus, reconstructing patio-temporally correlated climate from pollen involves estimating a complex, non-linear relationship from multinomial data making traditional Markov Chain Monte Carlo methods difficult. In this work, I apply a Polya-gamma data augmentation scheme to enable conjugate parameter updates and reduce computational costs, allowing for Bayesian paleoclimate reconstructions from pollen to be performed at regional-to-continental scales.

Analysis of wearable devices data in biomedical studies
Organizer: Ken Wang, Fred Hutch Cancer Research Center

Streamlining the collection and pre-processing of accelerometry data in large cohort studies and clinical trials
Jacek K. Urbanek, Phd, Meng (Johns Hopkins Medicine), Marta Karas (Johns Hopkins Bloomberg School of Public Health), Jennifer A. Schrack, Phd (Johns Hopkins Bloomberg School of Public Health)

Wearable physical activity (PA) monitors are small, noninvasive devices that can be worn on various body locations to measure human movement in a free-living environment. These measurements can be used to quantify multiple objective lifestyle and functional characteristics including volumes and fragmentation of physical activity, diurnal rhythms, sleep, and mobility. Following large, high-profile observational studies that successfully implemented wearable PA monitors, a growing number of health researchers strive to leverage wearable data across more focused clinical research projects. However, novelty, complexity, and the sheer size of wearable data create logistical and analytical challenges for first-time users, that may ultimately limit the appeal of this technology. We have developed a set of universal protocols, aids, and tools to help simplify and streamline the collection and processing of wearable data in health studies. These include unified experimental design, training materials, network and hardware infrastructure, as well as open-access statistical analysis software. We showcase proposed solutions in the existing clinical and intervention studies.

Graph-based tests on mean and variance components of the repeatedly assessed physical activity density objects
Haochang Shou (University of Pennsylvania)*

Repeated measures of wearable sensor data over multiple days have become commonly available in biomedical research and longitudinal studies. There is an increasing interest in modeling the continuous distribution of activity intensities. Such data have complex multivariate dependency structures and are often sampled from an arbitrary non-Euclidean metric space where the traditional testing methods might fail. We investigate the probability densities of daily physical activity measures as densely assessed by accelerometers. We propose a set of novel non-parametric graph-based test statistics to capture various possible alternatives in mean, inter- and intra-subject variability across clinical groups. Our proposed tests exhibit substantial power improvements over existing methods and are shown to preserve the type I errors under finite samples, as shown through simulation studies. The proposed tests are demonstrated to provide insights in the location, inter- and intra-subject variability of the daily physical activity distributions among participants with mood disorders as compared to healthy controls.

Functional data analysis methods for characterizing physical activity intensity and duration using accelerometry data
Chongzhi Di (Fred Hutchinson Cancer Research Center), Xu Wang (University of Washington), Guangxing Wang (Fred Hutchinson Cancer Research Center), Andrea Lacroix (University of California, San Diego)*

Accelerometers are widely used to objectively measure physical activity in large-scale epidemiological studies. It is of great scientific interest to explore patterns of activity intensity, duration and frequency and investigate whether dose-response relationships between activities accumulated from varying intensity and bout durations and health outcomes. We propose functional data analysis methods to flexibly model the distribution of bout lengths for each subject, which does not depend on threshold values or parametric distributional assumptions. Functional principal component analysis is adapted to provide low dimensional summary scores that characterize the distribution of bout durations for each subject, while functional regression estimate a dose-response relationship between bout durations and health outcomes. We apply these methods to the Objective Physical Activity and Cardiovascular Health (OPACH) Study, an ancillary study of the Women's Health Initiative, for quantifying associations between physical activity accumulation patterns and cardio-metabolic risk factors.

Eliciting longitudinal physical activity patterns using densely sampled accelerometry
Loki Natarajan (University of California San Diego), Wenyi Lin (University of California San Diego), Chongzhi Di (University of Washington Seattle)*

Accelerometers are widely used for tracking human movement and can provide minute-level (or even 30 Hz-level) estimates of physical activity (PA). Functional data methods, which model the entire accelerometer activity profiles as functional curves, are able to leverage the full information content of these densely sampled inputs. In this talk, we implement functional principal component analysis (FPCA) to study diurnal variation in accelerometer-derived longitudinal PA data. We then test the longitudinal associations between these patterns and health outcomes. The results show that the health outcomes are strongly associated with PA variation, at both the subject- and visit-level. In addition, we reveal that timing of PA during the day can impact outcomes, a finding that would not be possible with daily (or weekly) PA summaries. Thus, our findings imply that the use of longitudinal FPCA can elucidate temporal patterns of multiple levels of PA inputs.

Valid statistical approaches in non-randomized oncology study data analysis

Organizer: Sunhee Ro, BeiGene Pharmaceuticals USA

How to Make a "Relatively Fair" Comparison without a Randomized Controlled Trial
Zhiyue Huang (Roche), Shanmei Liao (Beigene), Yujie Zhong (Shanghai University of Finance and Economics)*

Randomized controlled trial (RCT) is considered as the golden standard to compare two therapies. Running a RCT is expensive and time-consuming. Inverse probability weighting (IPW) method can be used to balance the baseline characteristics and mimic a RCT. However, the individual level patients data of competitors is always unavailable. Under such circumstance, the matching-adjusted indirect comparison method (MAIC) and the simulated treatment comparison (STC) can be used to indirectly compare two therapies. In this talk, I am going to tell a fake story based on a real trial data. Through this story, I will introduce the STC, MAIC and IPW, as well as our proposed method Relaxed MAIC as a sensitivity analysis. I will compare the methods with a small simulation study. The class of indirect comparison methods (MAIC, STC and R-MAIC) could be used to determine model assumption at study design stage and are often required by reimbursement agencies for risk-benefit evaluation.

Propensity-score based vs regression based approach for adjusting bias in treatment effect estimate from non-randomized, cross-trial comparison
Sunhee Ro (Sierra Oncology)*

Propensity-score (PS) based methods have been increasingly popular for bias adjustment in the treatment effect (TE) estimation in non-randomized data. However, its difference from the traditional multiple regression is not well known. Sen et al (2007) laid out the mathematical framework of PS based stratification vs the multiple regression. The regression coefficient for the TE in the model is equivalent to the (average) TE only under the strongly ignorable treatment assignment. The paper also discusses when the conditional unbiasedness of PS stratification estimator is achieved. Biondi-Zoccai et al (2011) compared PS methods (stratification and adjustment as covariate) and multiple regression through extensive review of simulation results. In this presentation, I will summarize and compare the statistical mechanisms of these methods by which TE estimates are constructed, in order to understand the relative performance of the methods across different types of data under theoretical frameworks. The aim is to help audience to get better understanding of strength/limitation of each method selected.

Accounting for Patient Selection in the Interpretation of Single Arm Phase 2 trials
Eric Holmgren (Beigene Inc.)*

In single arm Phase 2 clinical trials where a large number of sites are expected to contribute less than a handful of patients to the study, a typical site will likely not enroll every eligible patient into the trial. Rather sites may have several studies open for enrollment at the same time and direct patients into the study that is viewed as most beneficial for them. This practice can skew the type of patients that are enrolled into the trial in ways that can not be observed simply by looking at patient characteristics. We will discuss this enrollment practice and the impact it can have on a phase 2 clinical trial. In addition, a framework is provided for interpreting study results that are affected by this enrollment practice.

A randomized Phase II design which brings in information on the control arm from past studies to reduce the sample size
Mithat Gonen (Memorial Sloan Kettering Cancer Center)*

Phase II cancer trials are traditionally designed as single-arm, with the null hypothesis chosen ahead of time, presumably from the literature. This approach is often criticized for its lack of

rigorous control but the practice continues, partly because a concurrent control arm will substantially increase the sample size. An unbalanced randomization with more patients randomized to the treatment arm is advocated as a way to keep the sample size in check, but when the resulting trial is analyzed using a traditional comparison, this planned imbalance actually reduces power. We propose a hybrid-approach where an imbalanced randomization scheme is implemented but the data for the control arm are augmented with historical information using a Bayesian approach. We will review the use of power priors and related methods to evaluate the gains from this procedure. While similar literature focuses on power improvements, we will also show implications for Type I error control.

Recent Statistical Developments in High-Dimensional Omics Sciences

Organizer: Debmalya Nandy, Postdoctoral Fellow, Department of Biostatistics & Informatics, Colorado School of Public Health at University of Colorado Anschutz Medical Campus

Identifying condition-specific patterns in large-scale genomic data

Hillary Koch (Penn State University), Cheryl Keller (Penn State University), Belinda Giardine (Penn State University), Guanjue Xiang (Penn State University), Ross Hardison (Penn State University), Qunhua Li (Penn State University)*

Joint analyses of genomic datasets obtained in multiple different conditions are essential for understanding the biological mechanism that drives tissue-specificity and cell differentiation. But it still remains computationally challenging even when the number of conditions is moderate. I will present CLIMB (Composite LIkelihood eMpirical Bayes), a statistical method which learns patterns of condition specificity present in genomic data by leveraging pairwise information. CLIMB provides a generic framework facilitating a host of downstream analyses, such as clustering genomic features sharing similar conditional-specific patterns and identifying which of these features are involved in cell fate commitment. It improves upon existing methods by boosting statistical power to identify biologically meaningful signals while retaining interpretability and computational tractability. We illustrate CLIMB on a ChIP-seq dataset and an RNA-seq dataset measured in hematopoietic lineages. These analyses demonstrate that CLIMB captures biologically relevant clusters in the data and improves upon commonly-used pairwise comparisons and unsupervised clusterings typical of genomic analyses.

Towards Mechanistic Inferences in Radiomics

Debashis Ghosh (Colorado School of Public Health, University of Colorado Anschutz Medical Campus), Emily Mastej (Computational Biosciences Program, University of Colorado Anschutz Medical Campus)*

Radiomics explores relationships between image-derived characteristics of a tumor and other parameters, including clinical outcomes and genomic profiles, including gene expression, somatic mutations, and DNA methylation. In spite of their state-of-the-art performance, use of these complex models comes at a cost. Because many of these classifiers are “black-box” in nature, clinicians consequently have a difficult time understanding the predictions. In this talk, we describe two complementary approaches towards more reasoned inferences in radiomics problems.

scClassify: multiscale classification of cells using single and multiple reference

Jean Yee Hwa Yang (The University of Sydney)*

Recent advances in large-scale single-cell transcriptome profiling have allowed for high-resolution studies of cell heterogeneity, developmental dynamics, and cellular communication across a wide range of biological systems. A key computational challenge in extracting meaningful information from these studies is accurate and automated cell type identification. To capitalise on

the large collection of well-annotated scRNA-seq datasets, we developed scClassify, a multiscale classification framework based on ensemble learning and cell type hierarchies constructed from single or multiple annotated datasets as references. scClassify enables the estimation of the sample size required for accurate classification of cell types in a cell type hierarchy and allows joint classification of cells when multiple references are available. We show that scClassify consistently outperforms existing classification methods and demonstrate its scalability on large single-cell atlases. Finally, we show how scClassify contributes to a generalizable workflow that explores cell-cell interactions in a set of single-cell COVID-19 datasets and assesses molecular patterns associated with disease severity.

Differential Expression Analysis using Kernel Machines for CyTOF data
Tusharkanti Ghosh (University of Colorado), Victor Lui (University of Colorado), Pratyaydipta Rudra (Oklahoma State University), Elena Hsieh (University of Colorado), Debashis Ghosh (University of Colorado)*

Mass Cytometry (CyTOF) is an advanced mass spectrometry technique that allows multiplexed measurements of protein expression at single-cell resolution. The emergence of CyTOF technology has given rise to the ability to observe changes in protein expression levels for high dimensional data. Current methodologies for analysis of differential expression (DE) borrow the principles from RNA sequencing data. Such principles are often not compatible with CyTOF data as it offers protein expression differences across multiple populations. Here, we propose cytoKernel, a novel differential expression (DE) method using a kernel-based score test for CyTOF data. We also present an unsupervised clustering algorithm as a pre-processing step before implementing the differential expression (DE) analysis where we identify similar phenotypic cell subpopulations as communities. We perform a systematic benchmark evaluation based on a subset of Common Variable ImmunoDeficiency (CVID) data and demonstrate that our method (cytoKernel) has a better performance than existing DE CytOF tools in terms of statistical power and false discovery rate control.

Recent developments in functional data analysis
Organizer: Chongzhi Di, Fred Hutchinson Cancer Research Center

Robust functional principal components for sparse longitudinal data
Matias Salibian-Barrera (The University of British Columbia), Graciela Boente (Universidad de Buenos Aires)*

We propose a new method for functional principal components analysis (FPCA) that can be applied to longitudinal data with few observations per trajectory. This method is robust against the presence of atypical observations in the data, and can be used naturally to derive a new non-robust FPCA approach for sparsely observed functional data. Our approach uses local regression to estimate the values of the covariance function taking advantage of the fact that for elliptically distributed random vectors the conditional location parameter of some of its components given others is a linear function of the conditioning set. This observation allows us to obtain robust FPCA estimators by using robust local regression methods. The finite sample performance of our proposal is illustrated through a simulation study that shows that, as expected, the robust method outperforms existing alternatives when the data are contaminated. Interestingly, we also see that for samples that do not contain outliers the non-robust version of our proposal compares favorably to the existing alternative in the literature.

Robust Functional Principal Component Analysis viaA Functional Pairwise Spatial Sign Operator
Ken Wang (Fred Hutchinson Cancer Research Center)*

In this talk, we present a new robust functional principal component analysis approach based on a functional pairwise spatial sign (PASS) operator, termed PASS FPCA. This includes robust estimation procedures for eigenfunctions and eigenvalues. Compared to existing robust FPCA approaches, the proposed PASS FPCA requires weaker distributional assumptions to conserve the eigenspace of the covariance function. The robustness of the PASS FPCA will be demonstrated via extensive simulation studies, especially its advantages in scenarios with asymmetric distributions. We will also look at an application to the accelerometry data from the Objective Physical Activity and Cardiovascular Health Study, which is a large-scale epidemiological study that investigates the relationship between objectively measured physical activity and cardiovascular health among older women.

Modeling trajectories using functional linear first-order differential equations
Julia Wrobel (Colorado School of Public Health), Jeff Goldsmith (Columbia University)*

We introduce a novel regression method that fuses concepts from functional linear regression and ordinary differential equations. Our work is motivated by novel data from an experiment exploring the relationship between neural firing rates and hand trajectories of mice performing a reaching task while under neurological assessment. This is an example from the increasingly common class of problems where outcome and responses are measured densely in parallel. For these data streams, we want to understand the relationship between inputs and outputs that are both functions measured on the same domain. Recent work using these data suggests that the dynamics of the arm during dexterous, voluntary movements are tightly coupled to neural control signals from the motor cortex. To better quantify how brain activity affects current and future paw position, our model incorporates initial position and has parameters that treat the relationship between the paw trajectory and the brain as a dynamical system of inputs and outputs, that state of which evolve over time. We compare our method to historical functional linear regression in simulations and on the mouse kinematic data.

Joint model for survival and multivariate sparse functional data with application to a study of Alzheimer's Disease
Li Cai (Yale University), Luo Xiao (North Carolina State University), Sheng Luo (Duke University)*

Studies of Alzheimer's disease (AD) often collect multiple longitudinal clinical outcomes, which are correlated and predictive of AD progression. It is of great scientific interest to investigate the association between the outcomes and time to AD onset. We model the multiple longitudinal outcomes as multivariate sparse functional data and propose a functional joint model linking multivariate func- tional data to event time data. In particular, we propose a multivariate functional mixed model to identify the shared progression pattern and outcome-specific progression patterns of the outcomes, which enables more interpretable mod- eling of associations between outcomes and AD onset. The proposed method is applied to the Alzheimer's Disease Neuroimaging Initiative study (ADNI) and the functional joint model sheds new light on inference of five longitudinal out- comes and their associations with AD onset. Simulation studies also confirm the validity of the proposed model. Data used in preparation of this article were obtained from the ADNI database.

Advances in statistical approaches for handling High-dimensional data
Organizer: Minjae Lee, University of Texas Southwestern

ConQuR-ing Batch Effects in Microbiome Profiling Studies using Conditional Quantile Mapping
Michael Wu (Fred Hutchinson Cancer Research Center), Wodan Ling (Fred Hutchinson Cancer Research Center)*

Batch effects in microbiome data are technical variation that result from differential handling of samples and can lead to both increased false positives and false negatives. We propose a strategy for batch correction based on regressing out the effect of batch while adjusting for additional covariates of interest. Specifically, we use a two-part quantile-regression based framework to regress the zero-inflated counts for each taxon on indicators for batch as well as additional covariates (including sequence depth). From the observed quantile function, we remove the effect of batch in order to obtain the batch corrected quantile function to which each observation is matched. Using a two-part quantile model makes minimal distributional assumptions while accommodating zero inflation, allowing for effective removal of batches while maintaining data structures. We illustrate our approach using several microbiome studies in which batch effects are inevitable due to large sample sizes.

Integrative Analysis of Multi-Omic Data via Sparse Multiple Co-Inertia Analysis
Eun Jeong Min (Catholic University of Korea), Qi Long (University of Pennsylvania)*

Multiple co-inertia analysis (mCIA) is a multivariate analysis method that can assess relationships and trends in multiple datasets. Recently it has been used for an integrative analysis of multiple high-dimensional -omics datasets. However, the estimated loading vectors from the existing mCIA method are non-sparse, which presents challenges for interpreting analysis results. We propose two new mCIA methods: 1) a sparse mCIA (smCIA) method that produces sparse loading estimates and 2) a structured sparse mCIA (ssmCIA) method that further enables the incorporation of structural information among variables such as those from functional genomics. The two proposed methods achieve simultaneous model estimation and feature selection and yield analysis results that are more interpretable than the existing mCIA. Our simulation studies demonstrate the superior performance of the smCIA and the ssmCIA methods compared to the existing mCIA. We also apply our methods to integrative analysis of transcriptomics data and proteomics data from a cancer study.

Extensions of machine learning methods for classification of objects based on high-dimensional measurements of embedded observations within each object.
Jose-Miguel Yamal (University of Texas Health Science Center at Houston), Frances Brito (University of Texas Health Science Center at Houston), E. Neely Atkinson (Rice University), Dennis D. Cox (Rice University)*

Multiple-instance learning typically involves classifying a macro-level object (e.g. a patient as having disease or not) based on measurements embedded within each object (e.g. microlevel observations like multiple images or cells with each patient). This data can be high dimensional with respect to the number of variables measured on each microlevel observation but also involves the complexity of multiple measurements per object that we want to classify. For example, the goal to classify a patient as having pre-cervical cancer or not based on high-dimensional measurements on a collection of around 2500 cells per patient. We propose extensions of some machine and statistical learning models to this structure of data and show, using simulations and real data, the benefit of using these methods over standard machine learning methods with macro-level summary features. These methods have broader applications in classification of hierarchical data including prediction of state-wide COVID-19 Rt based on data at the county level, hospitals using electronic medical records, schools using student measurements, and more.

A Landscape of Acquired Allelic Imbalance across the Cancer Continuum
Paul Scheet, Phd (The University of Texas MD Anderson Cancer Center)

Somatically-acquired chromosomal copy number alterations (SCNAs) are established factors in cancer initiation and have recently been implicated as a marker for cancer risk. In typical settings their detection becomes extremely challenged when the aberrant cell fraction (or tumor purity) is below 10-20%. Yet, this range may be critical for early detection and diagnostics since for such

applications the samples of interest will be comprised of heterogeneous mixtures of cells with a substantial component of DNA from normal (i.e. representing the germline) rather than aberrant (e.g. the tumor) sources. To address, we present a statistical technique that draws power from modeling the dependence among within-sample allele frequencies induced by SCNAs, leveraging the haplotype structure of the human genome. We demonstrate our method by elucidating large SCNAs, which would not otherwise be detected, in applications of cancer biology, implicate systematic errors in The Cancer Genome Atlas, and re-analyze 10 genome-wide association studies to reveal a 3-fold higher rate of SCNAs, supporting its use as a biomarker and yielding insights on background mutation patterns.

Analysis of health outcomes data with complex correlation structures

Organizer: Ann Lazar, University of California San Francisco, Associate Professor

Correlated gap time analysis with flexible hazards applied to pulmonary exacerbations in the EPIC Observational Study

Elizabeth Juarez-Colunga (University of Colorado Anschutz Medical Campus), John D Rice (University of Colorado Anschutz Medical Campus), Rachel L Johnson (University of Colorado Anschutz Medical Campus), Brandie D Wagner (University of Colorado Anschutz Medical Campus), Edith T. Zemanick (University of Colorado Anschutz Medical Campus), Margaret Rosenfeld (Seatle Children's Hospital)*

Cystic fibrosis clinical trials often use time to first pulmonary exacerbation (PEx) or total PEx count as endpoints. Use of these outcomes may fail to capture patterns or timing of multiple exacerbations. Analysis of gap times between PEx provides a useful framework to understand risks of subsequent events, particularly to assess if there is a temporary increase in hazard of a subsequent PEx following the occurrence of a PEx. This may be useful for estimating the amount of time needed to follow patients after a PEx, but requires the analysis to address the issue of correlation between gap times within a patient. We propose a smoothed hazard for gap times to account for elevated hazards after exacerbations, with cluster-robust standard errors estimated using either a sandwich covariance estimator or resampling procedures. A simulation study was conducted to explore model performance when misspecified and was able to appropriately estimate parameters in all situations with an underlying change-point. Models with different change-point structures and trends are compared using Early Pseudomonas Infection Control Observational Study (EPIC) data.

Use of copulas for analyzing discrete longitudinal and clustered data

N Rao Chaganty (Old Dominion University)*

Health sciences research and clinical trials often track treatments on individual subjects or families, resulting in longitudinal or familial data. Statistical analysis of such data is straight-forward when response variables are continuous, as the multivariate normal distribution can be used to model both potential predictors of the responses and also the dependence between repeated measurements or within families. However, the restricted ranges of some discrete observations makes longitudinal or clustered analyses challenging, as they lead to stringent constraints on some parameters in the multivariate probability distributions for these discrete outcomes. Copula functions provide a powerful alternative since they separate the dependence modeling from the marginal distributions, and hence any restrictions from the range of the outcome, alleviating some analytical difficulties posed by traditional methods. In this talk, I will give a short introduction to copulas and through motivating examples elucidate the use of these models in analyzing data that occur in health sciences and clinical trials.

The mixed model for repeated measures for cluster randomized trials

Melanie Bell (University of Arizona)*

Cluster randomized trials allocate intact clusters (hospitals, families, communities) to intervention and control rather than individuals Analysis of these trials must take into account correlation due to both clustering and repeated measures, if the trial is longitudinal. The mixed model for repeated measures (MMRM), is a popular choice for individually randomized trials with longitudinal continuous outcomes. This maximum likelihood-based model uses an unstructured time and covariance structure and its appeal is due to 1) avoidance of model misspecification and 2) its unbiasedness for data which are missing completely at random or missing at random. We extend the MMRM to cluster randomized trials and undertook a simulation experiment to test statistical properties. When simulating under the null, we found that type I error was nominal. When simulating a treatment effect we found that estimates were unbiased when data were complete and when data were missing at random.

High-dimensional inference with applications to -omics data

Organizer: Tusharkanti Ghosh, University of Colorado, Anschutz Medical Campus

Compositional Data Analysis using Kernels in Mass Cytometry Data

Pratyaydipta Rudra (OKlahoma State University), Ryan Baxter (University of Colorado Anschutz Medical Campus), Elena Hsieh (University of Colorado Anschutz Medical Campus), Debashis Ghosh (University of Colorado Anschutz Medical Campus)*

Cell type abundance data arising from mass cytometry experiments are compositional in nature. Classical association tests do not apply to the compositional data due to their non-Euclidean nature. Existing methods for analysis of cell type abundance data suffer from several limitations for high-dimensional mass cytometry data, especially when the sample size is small. We proposed a new multivariate statistical learning methodology, Compositional Data Analysis using Kernels (CODAK), based on the kernel distance covariance (KDC) framework to test the association of the cell type compositions with important predictors (categorical or continuous) such as disease status. CODAK scales well for high-dimensional data and provides satisfactory performance for small sample sizes. We conducted simulation studies to compare the performance of the method with existing methods of analyzing cell type abundance data from mass cytometry studies. The method is also applied to a high-dimensional dataset containing different subgroups of populations including Systemic Lupus Erythematosus (SLE) patients and healthy control subjects.

Mechanism-Aware Imputation: A two-step approach in handling missing values in metabolomics

Jonathan P. Dekermanjian (Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus), Elin Shaddox (Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus), Debmalya Nandy (Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus), Debashis Ghosh (Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus), Katerina Kechris (Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus)*

We propose a mechanism-aware imputation algorithm that estimates the underlying missingness pattern of a sparse dataset and then imputes values according to specific missingness mechanisms. Motivated by metabolomics data sets, our mechanism-aware imputation approach incorporates a random forest classifier to distinguish missing completely at random (MCAR) values from missing not at random (MNAR) values prior to imputation. We also demonstrate the performance of our approach compared to alternative imputation approaches through simulated datasets of varying patterns and levels of missingness in order to determine the best performing imputation algorithm for a particular missingness mechanism. Our model is applied to an

untargeted metabolomics study in plasma for chronic obstructive pulmonary disease (COPD), with the goal of reducing bias, which can improve statistical power for downstream analyses.

An Exploration of Multiple-Testing Correction Methods in Large-Scale Omics Studies
Debmalya Nandy (Presenter) (University of Colorado Anschutz Medical Campus), Debashis Ghosh (University of Colorado Anschutz Medical Campus), Katerina Kechris (University of Colorado Anschutz Medical Campus)*

High-throughput data are very prevalent in "-omics" sciences (e.g. Genomics, Metabolomics) containing measurements on several hundreds or thousands of variables. In tests of association of these "predictors" with a clinical outcome of interest, multiple-testing corrections mitigate false positives and false negatives among the statistically "significant" discoveries. Many corrections involve estimating the "effective" number of tests (number of statistically independent predictors among all original ones) and then using that for a Bonferroni-type adjustment to obtain the point-wise significance level (for a preset overall type-I error rate). Such practice is commonplace in Genome-Wide Association Studies (GWAS) but is also relevant to Metabolome-Wide Association Studies (MWAS). We review several multiple-testing procedures used for GWAS along with one recently published for MWAS, most of which are based on eigen-analysis of the Pearson's correlation matrix of the predictors. We propose using the distance correlation and examine its performance on some real metabolomics datasets. Keywords: Multiple comparisons; Correlated test statistics; Eigenvector; Family-wise error rate

CCmed: Cross-condition mediation analysis for identifying replicable trans-associations mediated by cis-gene expression
Fan Yang (University of Colorado Anschutz Medical Campus)*

Trans-acting expression quantitative trait loci (eQTLs) collectively explain a substantial proportion of expression variation, yet are challenging to detect and replicate since their effects are often individually weak. A large proportion of genetic effects on distal genes are mediated through cis-gene expression. We proposed a Cross-Condition Mediation analysis method (CCmed) for detecting cis-mediated trans-associations with replicable effects in relevant conditions/studies. CCmed integrates cis-association and gene-gene conditional correlation statistics from multiple tissues/studies. Motivated by the bimodal effect-sharing patterns of eQTLs, we proposed two variations of CCmed, CCmed_most and CCmed_spec for detecting cross-tissue and tissue-specific trans-associations, respectively. We analyzed data of 13 brain tissues from the Genotype-Tissue Expression (GTEx) project, and identified trios with cis-mediated trans-associations across brain tissues, many of which showed evidence of trans-association in two replication studies. We also identified trans-genes associated with schizophrenia loci in at least two brain tissues.

frontiers of statistical genomics: deep learning and beyond

Organizer: Wei Sun, Fred Hutchinson Cancer Research Center

A new clustering algorithm for assigning cells to known cell types according to marker genes
Hongyu Guo (University of Notre Dame), Jun Li (University of Notre Dame)*

On single-cell RNA-sequencing data, we consider the problem of assigning cells to known cell types, assuming that the identities of cell-type-specific marker genes are given but their exact expression levels are unavailable, that is, without using a reference dataset. Based on an observation that the expected over-expression of marker genes is often absent in a nonnegligible proportion of cells, we develop a new clustering method called scSorter. scSorter allows marker genes to express at a low level and borrows information from the expression of non-marker genes.

On both simulated and real data, scSorter shows much higher power compared to existing methods.

DeepGWAS to Enhance GWAS Signals for Neuropsychiatric Disorders via Deep Neural Network
Gang Li (Department of Statistics and Operations Research, University of North Carolina at Chapel Hill), Jia Wen (Department of Genetics, University of North Carolina at Chapel Hill), Patrick F Sullivan (Departments of Genetics and Psychiatry, University of North Carolina at Chapel Hill; 4. Department of Medical Epidemiology and Biostatistics and Department of (Clinical) Genetics, Karolinska Institutet), Yun Li (Departments of Genetics and Biostatistics)*

Increasing sample size for neuropsychiatric disorders is challenging. Thus, methods on existing samples via enhanced computational approaches are much needed. Here, we trained a 15-layer deep neural network, DeepGWAS, to enhance GWAS signals by leveraging summary statistics for schizophrenia (SCZ) together with brain related functional annotations. We then applied the trained model to enhance GWAS results from the 2nd largest SCZ study, and compared with those from the largest SCZ GWAS (serving as the working truth). Our enhanced GWAS improved sensitivity by 14% (from 75.6% to 89.2%). Our DeepGWAS outperformed alternative methods (logistic regression, random forest, and XGBoost), showing the highest precision rate (by 5.4-44.1%) with reduced false positives. We further applied DeepGWAS trained on SCZ GWAS to enhance Alzheimer's disease (AD) GWAS results. We observed similarly promising results, revealing 10 loci missed by input GWAS with 6 confirmed by latest larger sample size studies. Our study suggests DeepGWAS a powerful tool for genetic studies of complex traits, especially for neuropsychiatric disorders where increases in sample sizes are difficult to attain in practice.

Integrating GWAS and multi-omics QTL summary statistics to elucidate disease genetic mechanisms via a hierarchical low-rank model
Yihao Lu (University of Chicago), Jin Liu (Duke-NUS Medical School), Lin Chen (University of Chicago)*

In the post-GWAS era, evidences suggested that many of trait-associated SNPs affect complex traits/diseases through their effects on expression levels and other omics traits. Extensive evaluations of genetic effects on omics traits have revealed an abundance of quantitative trait loci for omics traits (omics QTLs). 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. 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.

Knockoff genotypes: value in counterfeit
Chiara Sabatti (Stanford), Matteo Sesia (University of Southern California), Emmanuel Candes (Stanford University)*

The framework of knockoffs has been recently proposed to perform variable selection under rigorous type-I error control, without relying on strong modeling assumptions. We extend the methodology of knockoffs to a rich family of problems where the distribution of the covariates can be described by a hidden Markov model. We develop an exact and efficient algorithm to sample knockoff variables in this setting and then argue that, combined with the existing selective framework, this provides a natural and powerful tool for performing principled inference in genome-wide association studies with guaranteed false discovery rate control. To handle the high level of dependence that can exist between SNPs in linkage disequilibrium, we propose a multi-resolution analysis, that simultaneously identifies loci of importance and provides results analogous to those

obtained in fine mapping. Appropriate models for haplotypes allow us to handle samples from multiple populations, and analysis of data from multiple environments brings us closer to the identification of causal variants. This is joint work with Matteo Sesia and Emmanuel Candes and other PhD students at Stanford University.

Statistical Considerations for N-of-1 Clinical Trial Designs

Organizer: Sonia Jain, University of California, San Diego

nof1: an R package for analyzing and presenting n-of-1 trials

Jiabei Yang (Brown University), Christopher Schmid (Brown University)*

N-of-1 trials, single participant trials with multiple treatments randomized over the study, help participants make treatment decisions through direct estimates of individual-specific treatment effect. Combining n-of-1 trials gives extra information for estimating population average treatment effect by repeatedly measuring participants compared with only once in randomized controlled trials. There is no standard tool for analyzing and presenting results for n-of-1 trials. We developed an R package, nof1, and implemented 1) Bayesian generalized linear models for individual n-of-1 trials and 2) Bayesian generalized linear mixed models for both the meta-analysis of a series of trials with a common set of treatments and the network meta-analysis of trials with different sets of treatments across individuals. The package provides summary statistics and plot options for presenting the data and the modeling results for both individual and a series of n-of-1 trials. We use nof1 to analyze the series of n-of-1 trials that evaluates the comparative effectiveness of different carbohydrate diets to usual diet on patients with inflammatory bowel disease.

A Bayesian-bandit adaptive design for N-of-1 clinical trials

Sama Shrestha (University of California San Diego), Sonia Jain (University of California San Diego)*

N-of-1 trials are randomized, multiperiod, crossover trials on a single subject. An aggregated N-of-1 design can be used to estimate the population effect from these individual trials. We present a Bayesian adaptive design for both the individual and aggregated N-of-1 trials using a multiarmed bandit framework that is estimated via efficient Markov chain Monte Carlo. A Bayesian hierarchical structure is used to jointly model the individual and population treatment effects. Our proposed adaptive trial design is based on Thompson sampling, which randomly allocates individuals to treatments based on the Bayesian posterior probability of each treatment being optimal. While we use a subject-specific treatment effect and Bayesian posterior probability estimates to determine an individual's treatment allocation, our hierarchical model facilitates these individual estimates to borrow strength from the population estimates. We present the design's performance via a simulation study and demonstrate that from a patient-centered perspective, subjects are likely to benefit from our adaptive design, in particular, for those individuals that deviate from the overall population effect.

Statistical considerations of Bayesian Model Parameters Under Fixed or Random Intercepts

Kexin Qu (Brown University), Christopher H Schmid (Brown University)*

Many meta-analytic models require a multilevel framework in which intercepts may be considered fixed or random. Choice of the type of intercept is controversial. Random intercepts facilitate predictions to new studies or new individuals in the case of N-of-1 trials. They also allow for borrowing of strength across individual units to improve predictions on units observed. But the resulting shrinkage of the intercepts can cause changes in the estimates of the treatment effects themselves. Because the number of parameters in a fixed intercept meta-analysis model is increasing at the same rate as the sample size, the maximum likelihood estimation assumption is violated. Previous research also suggests that although the fixed intercept model may bias the

between-study variance parameter and that the degree of bias might depend on the way the model is parameterized. Whether and how the parameters in a Bayesian model are also biased is currently unknown. This study investigates the sensitivity of Bayesian model parameters to different parameterizations of the multilevel model with fixed or random intercepts. It applies the findings to aggregated N-of-1 trial data and meta-analysis.

Modeling Individual Goal Achievement Behavior Using Bayesian Networks
Christian Pascual (University of California, San Diego)*

Prescribing daily activity goals can motivate individuals to be healthy and active, but achievement of these goals is influenced by both intrinsic and extrinsic factors. We used network analysis to try to identify what factors contribute to or against goal achievement in an N-of-1 study. We found that some individuals consistently achieve their goals independent of any outside factor, while others' achievement is influenced by their past effort or perceived business of their day. Identifying individuals in this latter group early in an N-of-1 trial can help researchers provide support to improve their daily activity and outcomes.

Novel methods in latent class analysis
Organizer: Sarah Schmiege, University of Colorado Anschutz Medical Campus

Latent Class Analysis with Time-Varying Covariate Effects: A Simulation Study and Empirical Example of LCA-TVEM
Bethany C. Bray (The University of Illinois at Chicago), John J. Dziak (The Pennsylvania State University), Stephanie T. Lanza (The Pennsylvania State University)*

Latent class analysis (LCA) summarizes multiple observed indicators in terms of an unobserved categorical variable, interpreted as dividing a heterogenous sample into homogenous subgroups. Time-varying effect modeling (TVEM) is another advanced method, allowing regression models in which the strength and/or direction of a relationship of a predictor to a response changes over time. We propose an approach called LCA-TVEM that combines the advantages of these methods, in which the probability of class membership can depend on a rich interaction between external covariates and a time variable. We demonstrate this approach by modeling differences in the prevalence of alcohol use patterns across ages for participants in a nationally representative dataset. We first identify the latent classes, then study how their relative prevalences depend nonlinearly on age, and then study how their relationship with gender can also depend on age. The proposed LCA-TVEM method is shown to facilitate flexibly asking and answering a new kind of question. Simulations are provided to show that the accuracy of parameter estimation under the assumed model is adequate given sufficient sample size.

Joint latent class modeling approach for predicting clinical outcomes with longitudinal profiles of biomarkers subject to limits of detection
Menghan Li (Penn State University), Ching-Wen Lee (Parexel International Co., Ltd.,), Lan Kong (Penn State University)*

In clinical and biomedical studies, identifying underlying subpopulations with different risk profiles is important to understand the complex mechanism of disease development and progression. Latent class modeling is one of the popular approaches to clustering subjects into different latent groups. In the context of biomarker studies where longitudinal biomarkers and clinical outcomes are measured, joint latent class models can be used to evaluate the value of biomarker trajectories in predicting a clinical endpoint. Our research focused on extending joint latent class models to accommodate biomarker measurements censored by the detection limits. We proposed a modified likelihood function and developed a Monte Carlo Expectation–Maximization (MCEM) algorithm for

model estimation based on Metropolis–Hastings method. We also described how to perform posterior classification, prediction of clinical outcomes, and test for conditional independence in the presence of censored biomarkers. The performance of our MCEM algorithm was examined in the simulation studies, and real data from a biomarker study were used to illustrate our methods.

Confirmatory latent class methods evaluating performance of threshold boundary and equality constraints
Sarah Schmiege (University of Colorado Anschutz Medical Campus)*

Most applications of latent class analysis (LCA) use an unrestricted approach to model estimation and class enumeration. Confirmatory methods for LCA, achieved via restrictions on model parameters during the estimation process, have utility for theory building and replication. A small number of applied and methodological studies have used parameter constraints, but have lacked standardization on type and performance of potential methods of constraint setting. The current study evaluated three methods of parameter constraint setting in LCA: (1) parameter equality constraints, (2) threshold boundary constraints, and (3) combined equality and boundary constraints. Using Monte Carlo simulations, all three methods were empirically compared to one another and to unconstrained models under population models that differed by parameter homogeneity and separation. Results evaluated parameter estimate recovery and model fit. Equality restrictions outperformed threshold boundary constraints, but were equivalent to or underperformed an unrestricted model parameterization under most conditions. Findings are discussed in terms of replication and validation of the class enumeration process.

Multilevel Latent Class Analysis for Cross-Classified Data Structures
Katherine Masyn (Georgia State University), Audrey Leroux (Georgia State University)*

Latent class analysis (LCA) has become increasingly widespread in the social and health sciences. However, despite the ubiquity of multilevel data in these same settings, multilevel modeling techniques for LCA are still in their infancy. Thus, it is no surprise that one of the more complicated multilevel modeling extensions involving cross-classified data structures (Beretvas, 2008), has never been applied to LCA. The purpose of this presentation is to demonstrate the specification and estimation of a cross-classified multilevel LCA model using data on students nested within schools and neighborhoods. We further explore, with an empirical example, the consequences of ignoring the cross-classified data structure for measurement and structural parameter estimates as well as for individual classification. Preliminary findings suggest that failure to account for the cross-classified data structure in a multilevel LCA produce biased variance estimates for the random multinomial intercepts, consistent with previous work on cross-classified modeling (Luo & Kwok, 2009; Meyers & Beretvas, 2006). Furthermore, latent class composition at the student level is distorted.

Statistical learning and inference in online, dynamic settings
Organizer: Jean Feng, University of California, San Francisco

Online Multiple Hypothesis Testing
Tijana Zrnic (University of California, Berkeley)*

In the online multiple hypothesis testing problem, the decision-maker is faced with a continual stream of hypothesis tests, with no a priori bound on the total number of tests, and they must make the decision of whether to reject any given hypothesis online — without any knowledge of the future tests. This setting increasingly characterizes large-scale testing schemes in science and industry, where new tests enter and old tests leave the platform in a possibly asynchronous, online fashion. In this talk, I will discuss methods for online multiple hypothesis testing that can ensure a pre-

specified bound on the false discovery rate at any point along the sequence of tests. I will also draw connections to more traditional, “offline” false discovery rate methods.

Bayesian logistic regression for online recalibration and revision of risk prediction models with guarantees

Jean Feng (University of California, San Francisco), Alexej Gossmann (U.S. Food and Drug Administration, CDRH/OSEL/DIDSR), Berkman Sahiner (U.S. Food and Drug Administration, CDRH/OSEL/DIDSR), Romain Pirracchio (University of California, San Francisco)*

After a clinical prediction model has been deployed, subsequently collected data can be used to fine-tune its predictions and adapt to temporal shifts. Because model updating introduces additional risks of over-fitting, such procedures must provide performance guarantees. To this end, we investigate two online logistic recalibration/revision procedures: Bayesian logistic regression (BLR) and dynamic Bayesian logistic regression with a Markov prior (MarBLR). We introduce notions of Type I and II regret to characterize the safety and effectiveness of these procedures and derive their respective bounds. In simulation studies, BLR and MarBLR recalibrated the original model to new patient populations, learned beneficial multivariable logistic revisions, and wrapped around black-box model-updating procedures to improve their safety. In a case study for predicting COPD diagnosis, the methods combined a gradient-boosted tree with a continually refitted version to steadily improve model calibration and discrimination. Both procedures consistently outperformed using a locked model and can improve the reliability of machine learning models over time.

Online non-parametric estimation and the quest for computational efficiency

Tianyu Zhang (University of Washington, Department of Biostatistics), Noah Simon (University of Washington, Department of Biostatistics)*

The optimal complexity of a predictive model (for balancing bias and variance) is dependent on the sample size. Thus, in the online scenario, it is natural to consider a model of growing complexity (as more observations are obtained). This leads to several questions: How do we appropriately increase complexity? Do we need to store and re-use old observations as complexity increases? How computationally efficient can we be without losing predictive performance? In this talk, I consider these questions in a simple model. I consider estimating a regression function that is smooth (lives in a Sobolev ellipsoid). I propose a simple stochastic gradient descent method (in a linear space of growing dimension) and show that 1) this estimator achieves the minimax L2 estimation error rate; and 2) Under some assumptions, this estimator uses the minimal memory of any estimator that achieves that minimax error rate. I relate this to more classical sieve/projection estimators, and to other stochastic gradient-based methods and ideas. These results also hold for estimating regression functions in an RKHS.

Online analysis of high-dimensional Gaussian graphical models

George Michailidis (U Florida)*

We present a novel scalable online algorithm for detecting an unknown number of abrupt changes in the inverse covariance matrix of sparse Gaussian graphical models with small delay. The proposed algorithm is based upon monitoring the conditional log-likelihood of all nodes in the network and can be extended to a large class of continuous and discrete graphical models. We also investigate asymptotic properties of our procedure under certain mild regularity conditions on the graph size, sparsity level, number of samples, and preand post-changes in the topology of the network. Numerical works on both synthetic and real data illustrate the good performance of the proposed methodology both in terms of computational and statistical efficiency across numerous experimental settings.

Recent advances in designs and quantitative analysis in immunological research

Organizer: Tao He, San Francisco State University

Characterization of the landscape of repertoire sequencing data with novel statistical approaches and advanced machine learning techniques

Li Zhang (University of California San Francisco), Hai Yang (University of California San Francisco), Zenghua Fan (University of California San Francisco), Tao He (San Francisco State University), Jason Cham (Scripps Green Hospital), Lawrence Fong (University of California San Francisco)*

T/B cells represent a crucial component of the adaptive immune system and have been shown to mediate anti-tumoral immunity and immune response to respiratory coronaviruses. Next generation sequencing of the T/B cell receptors is used to profile the T/B cell repertoire (Rep-seq). We develop a customized analysis pipeline to characterize and investigate the landscape of Rep-seq. We perform network analysis on Rep-seq data based on the sequence similarity, then quantify repertoire network by network properties and correlate with the clinical outcomes. In addition, we identify shared amino acid motifs across the entire repertoire and assess their relationship with the clinical outcomes by a novel ensemble feature selection approach. Furthermore, we introduce Bayes factor to incorporate clonal generation probability and real data abundance to identify the antigen-driven clones. The applications of the proposed approach in the cancer immunotherapy studies and SARS-CoV-2 patients show that network architecture of immune repertoire can reveal the mechanisms of the adaptive immune system responses to immunotherapies and the clinical outcome.

Alternative Analysis Methods for Non-proportional Hazards in Cancer Immunology Studies

Ray Lin (Genentech/Roche)*

Non-proportional hazards have been observed in cancer immunology studies, including the delayed treatment effect and the diminishing effect caused by the confounding due to follow-up therapies. The log-rank test loses power and the standard Cox model usually produces biased estimates under such conditions. These statistical impacts may provide misleading clinical interpretation of the results and may also lead to false-negative studies. The Non-Proportional Hazards Cross-Pharma Working Group was formed in 2016 with more than 15 companies across the pharmaceutical industry and have been conducting research to evaluate alternative analysis methods under non-proportional hazards. The findings and the recommendations of the Working Group will be presented and discussed.

Design for immuno-oncology clinical trials involving non-proportional hazards patterns

Zhenzhen Xu (FDA), Bin Zhu (NCI), Yongsoek Park (University of Pittsburg)

A typical challenge facing the design and analysis of immuno-oncology (IO) trials is the prevalence of nonproportional hazards (NPH) patterns manifested in Kaplan-Meier curves under time-to-event endpoints. The NPH patterns would violate the proportional hazards assumption, and yet conventional design and analysis strategies often ignore such a violation, resulting in underpowered or even falsely negative IO studies. In this article, we explore, both empirically and analytically, the fundamental causes for the occurrence of various NPH patterns and then present novel design and analysis strategies to properly address such issue. Empirical studies demonstrate that the proposed strategies can ensure adequate study power, whereas the conventional alternative leads to a severe power loss. More importantly, the proposed strategies pinpoint a solution to enhance the study efficiency, alleviate the NPH patterns and outline a path towards precision immunotherapy design.

Recent Development in Interrupted Time Series Methods

Organizer: Maricela Cruz, Kaiser Permanente Washington Health Research Institute; University of Washington Department of Biostatistics

Power and sample size calculation for interrupted time series analyses of count outcomes
Shangyuan Ye (Oregon Health and Science University)*

Interrupted time series (ITS) -- a quasi-experimental design -- is often used to evaluate the effectiveness of a health policy intervention. When the outcome of interest is rare, for example, for certain hospital-acquired infections, the common practice is to focus on count outcomes. However, analyzing ITS with count outcomes is challenging due to the need to consider possible overdispersion and the need to account for serial correlation. In this talk, I will discuss two time series models of count outcomes: the observation-driven model and the parameter-driven model. I will also introduce simulation-based approaches to calculate the sample size and power for ITS studies with aggregated count outcomes.

An Improved Method for Analysis of Interrupted Time Series (ITS) Data: Accounting for Patient and Healthcare Heterogeneity Using Weighted Analysis
Joycelyne Efua Ewusie (University of Ottawa), Joseph Beyene (McMaster University), Lehana Thabane (McMaster University), Sharon Straus (Li Ka Shing Knowledge Institute), Jemila Hamid (University of Ottawa)*

Interrupted time series (ITS) design involves collecting data across multiple time points before and after the implementation of an intervention to assess the effect of the intervention on an outcome. Segmented Regression is the most common statistical method used in ITS analysis. Despite it‚Äôs frequency of use, segmented regression when applied to aggregated data experience aggregation bias which results in imprecision of estimates and loss of power. Our main objective is to propose a weighted segmented regression method where the variability associated with aggregated data is accounted for in the analysis of ITS data. We also aim to statistically compare our proposed method with traditional segmented regression to assess model performance using well established criteria.

The two methods, weighted segmented regression and segmented regression are compared using data simulated from a normal distribution. The hierarchy of performances of the methods are assessed in terms of bias, mean squared error, level and power. Practical application of our proposed method is illustrated using real data.

Birds of a feather flock together: Comparing controlled pre-post designs
Carrie Fry (Vanderbilt University School of Medicine), Laura Hatfield (Harvard Medical School)*

This study formalizes comparative interrupted time series (CITS) using the potential outcomes framework; compares two version of CITS to two versions of difference-in-differences (DID); and re-analyzes three previously published papers using these models. Though written differently and preferred by different research communities, the general version of CITS and DID with group-specific pre-trends are the same: they yield the same counterfactuals and identify the same treatment effects. In a re-analysis with evidence of divergent pre-period trends, failing to account for this in standard DID led to an 84% smaller effect estimate than the more flexible models. In a second re-analysis with evidence of non-linear outcome trends, failing to account for this in linear CITS led to a 28% smaller effect estimate than the more flexible models. We recommend detailing a causal model for treatment selection and outcome generation and the required counterfactuals before choosing an analytical approach. The more flexible versions of DID and CITS can accommodate features often found in real data, namely, non-linearities and divergent pre-period outcomes trends.

A formal test for the existence of a change point in Interrupted Time Series

Maricela Cruz (Kaiser Permanente Washington Health Research Institute), Hernando Ombao (king abdullah university of science and technology), Daniel L Gillen (University of California Irvine)*

According to the 2018 Annual Review of Public Health, interrupted time series (ITS) designs may be the only feasible recourse for studying the impacts of large-scale public health policies. ITS designs borrow from case-crossover designs and function as quasi-experimental methodology able to retrospectively analyze the impact of an intervention and account for autocorrelation. Statistical models used to analyze ITS designs inherently assume a change point exists and either restrict the interruption's effect to a preset time point or remove data for which the intervention effects may not be fully realized. We propose the `supremum Wald test' (SWT), a test for the existence of a change point over a predetermined set of plausible change points. The SWT is implemented within ITS models that allow for inference regarding the estimation of a change point in the presence of a potentially lagged (or anticipatory) treatment effect. We provide empirical simulation studies to assess the type one error rate of the SWT and power for detecting specified change point alternatives. We also provide a brief overview of a toolbox that implements the SWT within an ITS model for continuous outcomes.

Invited Sessions Sponsored by IMS

Recent Advances in Neuroimaging Analysis

Organizer: Lexin Li, University of California, Berkeley

Time-varying l_0 optimization for Spike Inference from Multi-Trial Calcium Recordings

Tong Shen Et Al (Department of Statistics, University of California, Irvine), Zhaoxia Yu (Department of Statistics, University of California, Irvine)*

Optical imaging of genetically encoded calcium indicators is a powerful tool to record the activity of a large number of neurons simultaneously over a long period of time from freely behaving animals. However, determining the exact time at which a neuron spikes and estimating the underlying firing rate from calcium fluorescence data remains challenging, especially for calcium imaging data obtained from a longitudinal study. We propose a multi-trial time-varying ℓ_0 penalized method to jointly detect spikes and estimate firing rates by robustly integrating evolving neural dynamics across trials. Our simulation study shows that the proposed method performs well in both spike detection and firing rate estimation. We demonstrate the usefulness of our method on calcium fluorescence trace data from two studies, with the first study showing differential firing rate functions between two behaviors and the second study showing evolving firing rate function across trials due to learning.

Brain connectivity-informed regularization methods in multi-modal imaging

Jaroslaw Harezlak (Indiana University), Damian Brzyski (Wroclaw University of Science and Technology), Timothy Randolph (Fred Hutchinson Cancer Research Center), Joaquin Goni (Purdue University), Kewin Paczek (Jagiellonian University), Aleksandra Steiner (University of Wroclaw)*

We address the problem of adaptive incorporation of multi-modal brain imagining data sources in multiple linear regression setting. In the presented example, we model scalar outcome dependence on the brain cortical properties, e.g. cortical thickness. We utilize connectivity and spatial proximity information to build an adaptive penalty terms in the regularized regression problem. The general idea of incorporating external information in regularization approach via linear mixed model representation has been recently established in our prior work: ridgified Partially Empirical Eigenvectors for Regression (riPEER). Here, we incorporate multiple sources of information, including structural and functional connectivity network structure as well as geodesic distance between the cortical regions, and estimate the regression parameters with multiple penalty terms via a riPEER extension called AIMER (Adaptive Information Merging Estimator for Regression). We present a simulation study testing various realistic scenarios and apply AIMER to data arising from the Human Connectome Project (HCP) study.

A Bayesian approach to joint modeling of matrix-valued imaging data and treatment outcome with applications to depression studies

Bei Jiang (University of Alberta), Eva Petkova (New York University), Thaddeus Tarpey (New York University), R. Todd Ogden (Columbia University)*

In this paper, we propose a unified Bayesian joint modeling framework for studying association between a binary treatment outcome and a baseline matrix-valued predictor. This framework establishes a theoretical relationship between the outcome and the matrix-valued predictor, although the predictor is not explicitly expressed in the model. Simulation studies are provided showing that the proposed method is superior or competitive to other methods, such as a two-stage approach and a classical principal component regression in terms of both prediction accuracy and estimation of association; its advantage is most notable when the sample size is small and the dimensionality in the imaging covariate is large. Finally, our proposed joint modeling approach is shown to be a very promising tool in an application exploring the association between baseline

electroencephalography data and a favorable response to treatment in a depression treatment study by achieving a substantial improvement in prediction accuracy in comparison to competing methods.

Functional Response Quantile Regression Model
Xingcai Zhou (Nanjing Audit University), Dehan Kong (University of Toronto), Adam Kashlak (University of Alberta), Rohana Karunamuni (University of Alberta), Linglong Kong (University of Alberta)*

In this paper, we propose a new functional response quantile regression model. A data driven estimation procedure is developed to estimate the quantile regression processes based on local linear approximation. Theoretically, we obtain the global uniform Bahadur representation of the estimator with respect to the time/location and the quantile level, and show that the estimator converges weakly to a two-parameter continuous Gaussian process. We then derive the asymptotic bias and mean integrated squared error of smoothed individual functions and their uniform convergence rates under given quantile levels. Based on the theoretical results, we introduce a global test for the coefficient functions and discuss how to construct simultaneous confidence bands. We evaluate our method through simulations and two applications from diffusion tensor imaging data and ADHD-200 functional magnetic resonance imaging data.

Topics in Causal Inference
Organizer: Lexin Li, University of California, Berkeley

Inference for algorithm-agnostic variable importance
Brian Williamson (Fred Hutchinson Cancer Research Center), Peter B. Gilbert (Fred Hutchinson Cancer Research Center), Noah Simon (University of Washington), Marco Carone (University of Washington)*

In many applications, it is of interest to assess the relative contribution of features (or subsets of features) toward the goal of predicting a response, that is, to gauge the variable importance of features. Most recent work on variable importance assessment has focused on describing the importance of features within the confines of a given prediction algorithm. However, such an assessment does not necessarily characterize the prediction potential of features and may provide a misleading reflection of the intrinsic value of these features. To address this limitation, we propose a general framework for nonparametric inference on interpretable algorithm-agnostic variable importance. We define variable importance as a population-level contrast between the oracle predictiveness of all available features versus all features except those under consideration. We then propose a nonparametric efficient estimation procedure that allows the construction of valid confidence intervals and tests, even when machine learning techniques are used. We discuss several examples, including a causally-motivated measure of variable importance based on individualized treatment rules.

Inference on Heterogeneous Quantile Treatment Effects via Rank-Score Balancing
Jingshen Wang (UC Berkeley)*

Understanding treatment effect heterogeneity in observational studies is of great practical importance to many scientific fields because the same treatment may affect different individuals differently. Quantile regression provides a natural framework for modeling such heterogeneity. In this paper, we propose a new method for inference on heterogeneous quantile treatment effects that incorporates high-dimensional covariates. Our estimator combines a debiased l1-penalized regression adjustment with a quantile-specific covariate balancing scheme. We present a comprehensive study of the theoretical properties of this estimator, including weak convergence of the heterogeneous quantile treatment effect process to the sum of two independent, centered Gaussian processes. We illustrate the finite-sample performance of our approach through Monte

Carlo experiments and an empirical example, dealing with the differential effect of mothers' education on infant birth weights.

Causal Estimation in Observational Data Subject to Missing by A Machine Learning Approach
Xiaochun Li (Indiana University)*

Observational medical databases increasingly find uses for comparative effectiveness and safety research. However, the lack of analytic methods that simultaneously handle the issues of missing data and confounding bias along with the onus of model specification, limit the use of these valuable data sources. We derived a novel machine-learning approach based on trees to estimate the average treatment effect. In order to evaluate causal estimation by model-free machine-learning methods in data with incomplete observations, we conducted a simulation study with data generated from known models of exposure, outcome and missing mechanisms. Thus the true causal effect was known and used as the benchmark for evaluations. Two settings were studied. We compared the bias and standard error of causal estimates from our method to a multiply robust parametric method, the complete case analysis (CC) and a regression analysis after multiple imputations (MI). The proposed methods were applied to a real observational data set of implantable cardioverter defibrillator use.

Invited Sessions Sponsored by JR

New directions in radiation epidemiology

Organizer: Munechika Misumi, Department of Statistics, Radiation Effects Research Foundation

Radiation risk estimation and statistical methods for the long-term follow-up studies of Japanese Atomic Bomb Survivors

Munechika Misumi (Department of Statistics, Radiation Effects Research Foundation)*

The Life Span Study (LSS) of the Radiation Effects Research Foundation is a long-term follow-up study of Japanese atomic bomb survivors. The result of LSS is an important resource for quantifying the risk of long-term effects of human exposure to ionizing radiation. Radiation epidemiology is a multidisciplinary field involving epidemiologists, statisticians, physicians, and radiation physicists and dosimetrists that has developed and applied sophisticated methods to estimate these risks. The application of these methods has resulted in the ability to estimate risks of radiation exposure to individual organs and to assess the accuracy and precision of these estimates in light of uncertainties in dose estimates. Traditionally, excess risk models are used incorporating corrections for dose estimate uncertainties. In this talk, the epidemiological studies and statistical methods will be introduced along with examples and discussion of possible future directions in this research.

Biologically based models of cancer risk in radiation research

Jan Christian Kaiser (Helmholtz Zentrum Muenchen)*

Biologically based models of cancer risk in radiation research The pioneering studies of the radiation risk in radioepidemiological cohorts were based on descriptive models which link excess rates of cancer incidence and mortality to radiation exposure by statistical association. To estimate the number of sporadic and radiation-induced cases descriptive models rely on convenient functional forms of dose responses and directly describe age-risk patterns. Against it, biologically-based models of cancer risk (BBCR models) models facilitate a more comprehensive consideration of biological processes for risk assessment. Compared to descriptive models dose responses are predicated on biological processes with indirect influence on risk. BBCR models improve understanding of radiation-related carcinogenesis by integrating molecular biology with epidemiology. BBCR models can harness information compiled for adverse outcome pathways for more accurate risk estimation.

Statistical issues in estimating factors affecting the individual response to radiation

Kyoji Furukawa (Kurume University)*

Stochastic effects of radiation exposure are variable between individuals, depending on possibly various factors including sex, age, environmental, genetics and epigenetic factors. Identification of radiosensitive subgroups is one of the primary interests for radiation protection. In practice, however, due to the lack of statistical power or unavailability of information, epidemiological risk evaluations often end up omitting potentially influential but statistically insignificant or unmeasured factors, which introduces more or less individual variations unaccounted in analysis. While such unobserved heterogeneity (frailty) can complicate the risk estimation at individual level through selection of less and less frail individuals over time, it has been little considered in radiation risk assessments. This study aims to evaluate the potential impact of plausible frailty variations by introducing random errors in an epidemiological risk analysis. A frailty variation of moderate size can explain the age-dependent risk pattern observed at cohort level equally well as with a conventional risk modification by age, while providing an individual interpretation of the estimated risk.

An application of multiple indicators, multiple causes measurement error models to adjust for dose error in RERF data
Carmen D. Tekwe (Indian University - Bloomington), Randy L. Carter (University at Buffalo)

Multiple Indicators, Multiple Causes (MIMIC) models are used by researchers to study the effects of an unobservable latent variable on a set of outcomes, when causes of the latent variable are observed. There are times however when the causes of the latent variable are not observed because measurements of the causal variable are contaminated by measurement error. We present (1) an extension of the classical linear MIMIC model to allow both Berkson and classical measurement errors, defining the MIMIC measurement error (MIMIC ME) model; (2) likelihood based estimation methods using the EM algorithm with Monte Carlo approximation to the integral in the E-step for the MIMIC ME model; and (3) obtain data driven estimates of the variance of the classical measurement error associated with log(DS02), an estimate of the amount of radiation dose received by atomic bomb survivors at the time of their exposure. The model was applied to study the effects of dyslipidemia, a latent construct and the effect of true radiation dose on the physical manifestations of dyslipidemia among Adult Health Study AHS cohort of atomic bomb survivors.

Advances in ecological data modelling
Organizer: Hideyasu Shimadzu, Loughborough University, UK

Integrating multiple sources of ecological data to estimate species abundance of woody plants at geographic scales
Keiichi Fukaya (National Institute for Environmental Studies), Buntarou Kusumoto (Kyushu University), Takayuki Shiono (University of the Ryukyus), Junichi Fujinuma (University of Tartu), Yasuhiro Kubota (University of the Ryukyus)*

The pattern of species abundance, represented by the number of individuals per species within an ecological community, is one of the fundamental characteristics of biodiversity. However, despite their obvious significance in ecology and biogeography, there is still no clear understanding of these patterns at large spatial scales. Here, we introduce a hierarchical modeling approach to estimate macroscale patterns of species abundance. Given the expense associated with collecting individual count data at the community level, we integrate spatially replicated multispecies detection-nondetection observations (specifically, vegetation surveys) and information on the geographical distribution of species to estimate species density at geographic scales. Using this approach, estimates of absolute abundance of 1248 woody plant species at a 10-km-grid-square resolution over East Asian islands across subtropical to temperate biomes are obtained. Results highlight the potential of the elucidation of macroscale species abundance that has thus far been an inaccessible but critical property of biodiversity.

Estimating Abundance from Animal Traces
Rafael Moral (Maynooth University), Iuri Ferreira (Federal University São Carlos), Niamh Mimnagh (Maynooth University), Luciano Verdade (University of São Paulo)

In this talk, we present a new method to estimate animal abundance based on the number of traces (e.g. footprints, fur, scats) found in a survey area. This method is suitable for both solitary and gregarious animals. We developed a fully Bayesian framework to estimate the number of groups, population density and the total number of traces found in the environment in a given moment, from the survey data and by considering a hierarchical Poisson model. We assumed prior knowledge about the surveyed tracks coverage (% of the total area) and the mean group size (ind./group). We were able to estimate the trace production rate and total abundance in the survey area, alongside the 95% credible intervals. However, when we assumed the trace production rate

as known, the credible intervals were much narrower. We studied different methods for estimating the full model, analyzing the implications from the assumption of incorrect traces production rates. We illustrate our approach with real survey data from a Eucalyptus farm located in Southeast Brazil.

Spatiotemporal modeling of an estuarine decapod using Bayesian inference: environmental drivers of juvenile blue crab abundance

A. Challen Hyman (William & Mary's Virginia Institute of Marine Science), Grace S. Chiu (William & Mary's Virginia Institute of Marine Science), Romuald N. Lipcius (William & Mary's Virginia Institute of Marine Science), Mary C. Fabrizio (William & Mary's Virginia Institute of Marine Science)

Nursery grounds substantially enhance secondary production of commercially exploited fish and crustacea populations by providing food and refugia for their juveniles. Previous small-scale studies for blue crabs have emphasized seagrass meadows as highly productive nurseries. Yet, to generalize inference of nursery function, identify highly productive regions, and inform regional management, it is vital to unify digitized data on structurally complex habitats with survey data over larger spatiotemporal scales. Thus, we construct five Bayesian hierarchical models with various spatiotemporal dependence structures on 22 years of data across temperate estuaries in Virginia to infer nursery habitat value for blue crabs. Our results indicate that 1) the nonseparable spatiotemporal model outperformed the simpler models in cross validations, and 2) salt marsh surface area and turbidity, not seagrass, are the strongest determinants of local juvenile blue crab production. These highlight the need to consider nursery function at multiple spatiotemporal resolutions, and therefore, spatiotemporal dependence in large scale fisheries catch data, in order for robust inference on local productivity.

Categorical data analysis to investigate spatial and temporal trend for Integrated Ecosystem Assessment in the Norwegian Sea

Hiroko Solvang (Institute of Marine Research)*

Integrated Ecosystem Assessment (IEA) is a set of approaches used for organizing scientific information at multiple scales, ecosystem components and across sectors to support marine ecosystem-based management. The integrated trend analysis in IEA is a method to summarize changes that have occurred in recent decades in the North Atlantic ecosystem and to highlight the possible connections among the physical, biological, and human ecosystem components. This method covers graphical analyses as well as univariate and multivariate statistical analyses. The observations used in the trend analysis are sometimes too short or too sparse for applying a statistical model. In such case, we find it practical to transfer the observations to categorical data like presence/absence at each observed location and explore independent or dependent tendencies among co-occurrent species at each location during the period of observation. To support our argument, we provide an approach based on the study by Sakamoto and Akaike 1998.

Invited Session Sponsored by CWS

The 200th Birth Anniversary of Florence Nightingale: Celebrating Women in Statistics - Past, Present, and Future.

Organizer: Nusrat Jahan, James Madison University

Empowering women in statistics for 50 years: History of the Caucus for Women in Statistics
Motomi Mori (St. Jude Children's Research Hospital), Jessica Kohlschmidt (Ohio State University), Wendy Lou (University of Toronto), Amanda Golbeck (University of Arkansas for Medical Sciences)*

To carry on the legacy of Florence Nightingale and to empower women statisticians, the Caucus for Women in Statistics (CWS) has been serving as an advocacy organization for women in statistics for the last fifty years. The CWS is an international, professional statistical society formed in 1971, for the education, employment and advancement of women in statistics. Its membership is open to anyone who supports CWS's mission and vision, from academia, industry, government and elsewhere. In this presentation. I will present fifty-year history of CWS, notable accomplishments by each decade, celebratory events scheduled this year, and future challenges for the CWS and women in statistics around the world.

Florence Nightingale Day: Inspiring and Passing the 'Lamp' to the Next Generation Statisticians
Shili Lin (Ohio State University)*

In 2018, the American Statistical Association and the Caucus for Women in Statistics jointly launched the annual Florence Nightingale Day (FN Day) in celebration of the event's namesake and all women in statistics. Our vision is to inspire a diverse group of students, especially women and young people from disadvantaged backgrounds to pursue a career in statistics and data science and to become future leaders in these fields. The specific goals for the day of the event are to expose upper middle and high school students, and especially girls, to statistics as a field of study and career option, with hands-on activities, panel discussions, and data challenges. Since 2018, the FN Day event has been (and will be) offered in a number of locations around the country, including Athens, Georgia; Chapel Hill, North Carolina; Columbus, Ohio; Washington DC (joined with the Korean International Statistical Society); and Dallas, Texas (October 2021). In this talk, I will describe our vision for expanding the event to other sites in the US and internationally, with the goal of encouraging everyone to consider hosting the event in your local community.

Paving the Way: Women as Mentors and Advocates for Junior Statisticians
Jessica Minnier (Oregon Health & Science University)*

Florence Nightingale is one of the most famous female statisticians in history, and as such is an inspiration for those interested in mathematics and statistics as a career. It is important for young and/or junior people to see representation in their field of people of all gender identities, orientations, race, ethnicity, and backgrounds. In this talk honoring Nightingale's contributions to science and the women's movement, I will recount some personal experiences with female mentors, highlight other important role models and advocates for diversity and representation in our field, and will address recent controversy regarding women as mentors in science.

Statistics Education in the field of Health Sciences
Nusrat Jahan (James Madison University)*

In 1856 serving as a nurse in the Crimean War, Florence Nightingale realized the power of data. In the middle of the Crimean War, she saved lives as well as collected pertinent information. Her impressive use of data to advocate health care policies was phenomenal. Her work highlighted the importance of statistics in the field of health sciences and medicine. Especially now in the era of big

data, data-driven decision-making processes have transformed the field of health sciences. This created a demand for statistics education in the health sciences sector. In this talk, we outline various ways of incorporating statistics education and research experience in the traditional undergraduate health sciences curriculum. We will also discuss the benefits of interdisciplinary collaborative projects with big data applications for students.

Contributed Sessions

Contributed Session 1

Simulating Bugs Over Time: A User-Friendly Guide to Simulating Longitudinal OTU Counts Using the Dirichlet-Multinomial Distribution

Nicholas Weaver (Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO), Brandie Wagner (Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO), Audrey Hendricks (Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO ; Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO)*

Recently, the scientific community has made great strides in studying the relationship between the microbiome and human health. Statistical tools and methods have been developed to identify the role of microbial communities in the gut, lungs, and skin. Most method development and comparisons have been performed cross-sectionally using simulated and real-life data. There is less known about the optimal method for studying the microbiome longitudinally. Additionally, simulation methods for realistic changes in OTU counts over time are lacking. Here we use the Dirichlet-Multinomial distribution to simulate baseline OTU counts and then perturb the baseline values to construct OTU counts over time. We find the Dirichlet-Multinomial distribution mirrors real data for moderate or large OTU relative abundance but will often miss small relative abundance OTUs. We further show that baseline counts can be perturbed overtime for differing levels of within-subject variability matching real life scenarios. Additionally, we create a publicly available and detailed RMarkdown vignette including visualizations for assessment so that others can easily simulate realistic longitudinal microbiome data.

Extension of the Condition-adaptive Fused Graphical Lasso and Application to Modeling Brain Region Co-Expression Networks

Souvik Seal (Department of Biostatistics and Informatics, CU Anschutz Medical Campus), Laura Saba (University of Colorado), Katerina Kechris (Department of Biostatistics and Informatics, CU Anschutz Medical Campus)*

Gaussian graphical modeling has been a standard tool for constructing gene co-expression networks. Harnessing shared information about related network structures across multiple different conditions (e.g., tumor and normal tissue) can significantly increase the power of analysis. In addition, examining condition-specific patterns of co-expression can provide insights into the underlying cellular processes activated in a particular condition. Condition-adaptive fused graphical lasso (CFGL) incorporates condition specificity in the estimation of multiple co-expression networks. However, the current implementation of CFGL in R can only accommodate three different conditions and is prohibitively slow for a moderate number of genes. We have developed a C/Python based package, rapid condition-adaptive fused graphical lasso (RCFGL), that can handle more than three conditions and is computationally more feasible. We have applied RCFGL to examine gene expression data from five different brain regions using a collection of stock rats. In summary, we have generalized the CFGL algorithm for more than three conditions overcoming its severe computational limitations.

The Impact of Continuity Corrections on Rare-Event Meta-Analysis

Brinley N Zabriskie (Brigham Young University), Jake Baldauf (Brigham Young University), Nolan Cole (Brigham Young University)*

Meta-analyses have become the gold standard for synthesizing evidence from multiple clinical trials. They are useful when clinical trials are small and the outcome is rare or adverse since individual trials often lack sufficient power to detect a treatment effect. However, when zero events are observed in one or both treatment arms in a trial, commonly used meta-analysis methods fail.

Continuity corrections, numerical adjustments to the data to make computation feasible, have been proposed to ameliorate this issue, but the impact of the various available continuity corrections on meta-analyses with rare events has not been explored. We compare several continuity corrections via a simulation study with a variety of commonly used meta-analysis methods. We consider how these continuity corrections impact important meta-analysis results, such as the estimated overall treatment effect, estimated heterogeneity variance, and Type I error rate.

Extension of the Two-Step Approach for Informative Dropout in Survival Analysis
Cristina Murray-Krezan (Division of Epidemiology, Biostatistics and Preventive Medicine, Department of Internal Medicine, University of New Mexico), V. Shane Pankratz (Division of Epidemiology, Biostatistics and Preventive Medicine, Department of Internal Medicine, University of New Mexico)*

The magnitude of risk for a time-to-event endpoint associated with a covariate of interest is frequently estimated from cohort studies by the use of survival analysis techniques. Longitudinal studies are widely known to suffer from dropout and missing data, and if dropout is not at random such estimates of association can be biased. We present an extension to the Two-Step Approach for nonignorable missing outcomes originally proposed for linear mixed models that we have extended to a survival analytic endpoint. This Extended Two-Step approach can provide a test for informative dropout and adjust measures of relative risk to reduce bias. This method first models the time to dropout with a frailty model to estimate the random effects associated with dropout. Next, these estimates are included as a covariate in a second frailty model for the outcome of interest to obtain relative risk estimates adjusted for unobserved variability due to potentially informative dropout. Through simulations and an application to decreased kidney function in children enrolled in the CKiD study, we demonstrate the benefits of this method, compared to the standard approach that treats dropout as ignorable.

Salmon stock forecasting using remote sensing covariates.
Mehnaz Jahid (University of Victoria), Maycira Costa (University of Victoria), Saman Muthukumarana (University of Manitoba), Wendell Challenger (LGL Limited), Laura Cowen (University of Victoria)*

The objective of this study is to understand how the ocean conditions data (sea surface temperature (SST) and chlorophyll-a (chl-a)) extracted from remote sensing satellites help to forecast the stock recruitment of Pacific salmon. To achieve this, we used spawner and recruitment data of Sockeye salmon from 1948-2016 of Pitt river stock, British Columbia. The remote sensing data (from 2003 to 2016) was extracted from the central and northern region of the Strait of Georgia, BC. SST from light-station data (from 1948-2016 of Entrance island and Pine Island) were also used for comparison. Ricker, Power and Larkin spawner-recruit models were used for the forecasting. Considering that the data might show temporal autocorrelation, an Auto Regressive-1 (AR(1)) component as also considered in all the models. We found that the Ricker AR(1) model with number of spawners as an explanatory variable explained the recruitment variation the best. However, since the remote sensing data does not have information from 1948- 2002, we used the data from 2003-2016 to run all the models and found that the Ricker model with SST and chl-a explained the variation the best.

A unified standardized selection probability to locate rare variants associated with highly correlated multiple phenotypes
Xianglong Liang (Department of Statistics, Pusan National University, Korea), Hokeun Sun (Department of Statistics, Pusan National University, Korea)*

In the past few decades, many statistical methods have been developed to identify rare variants associated with a complex trait or a disease. Recently, rare variant association studies with multiple phenotypes have drawn a lot of attentions because association signals can be boosted when rare variants are related with more than one phenotype. Most of existing statistical methods to identify rare variants associated with multiple phenotypes are based on a group test, where a gene or a

genetic region is tested one at a time. However, these methods are not designed to locate individual rare variants within a gene or a genetic region. In this article, we propose a unified standardized selection probability to locate individual causal rare variants associated with highly correlated multiple phenotypes. In our simulation study, we demonstrated that the propose method outperforms the existing selection methods in terms of true positive rate, when phenotype outcomes are highly correlated with each other. We also applied the proposed method to our wild bean data set that consists of 10,783 rare variants and 13 correlated amino acids.

New selection method to identify pleiotropic variants associated with both quantitative and qualitative traits
Kipoong Kim (Pusan National University), Hokeun Sun (Pusan National University)*

In recent genetic association studies, statistical methods to identify pleiotropic variants associated with multiple phenotypic traits have been developed, since susceptible variants with small effects could be easily missed in association studies based on a single trait. However, most of the existing methods to identify pleiotropic variants are designed for only quantitative traits even though pleiotropic variants are often associated with both quantitative and qualitative traits. There are some meta-analysis methods which basically integrate summary statistics of individual variants associated with either a quantitative or qualitative trait. But, these methods cannot account for correlations between genetic variants. In this article, we propose new selection probability computation to prioritize individual variants associated with both quantitative and qualitative traits. For each phenotypic trait, coefficients of elastic-net regularization are first estimated and then they are additively combined to compute selection probability of individual variants. We demonstrated that the proposed methods outperform the existing methods in both simulation studies and real data analysis.

Contributed Session 2

Improving Random Forest Predictions in Small Datasets from Two-phase Sampling Designs
Sunwoo Han (Fred Hutchinson Cancer Research Center), Brian D. Williamson (Fred Hutchinson Cancer Research Center), Youyi Fong (Fred Hutchinson Cancer Research Center)*

While random forests are one of the most successful machine learning methods, it is necessary to optimize their performance for use with datasets resulting from a two-phase sampling design with a small number of cases – a common situation in biomedical studies, which often have rare outcomes and covariates whose measurement is resource-intensive. Using an immunologic marker dataset from a phase III HIV vaccine efficacy trial, we seek to optimize random forest prediction performance using variable screening, class balancing, weighting, and hyperparameter tuning. We further show that prediction performance on this dataset can be improved by stacking random forests and generalized linear models trained on different subsets of predictors, and that the extent of improvement depends critically on the dissimilarities between candidate learner predictions.

A High-dimensional Mediation Model for a Neuroimaging Mediator: Integrating Clinical, Neuroimaging, and Neurocognitive Data to Mitigate Late Effects in Pediatric Cancer
Jade Xiaoqing Wang (St. Jude Children's Research Hospital), Yimei Li (St. Jude Children's Research Hospital), Wilburn E. Reddick(St. Jude Children's Research Hospital), Heather M. Conklin(St. Jude Children's Research Hospital), John O. Glass(St. Jude Children's Research Hospital), Arzu Onar-Thomas(St. Jude Children's Research Hospital), Amar Gajjar(St. Jude Children's Research Hospital), Cheng Cheng (St. Jude Children's Research Hospital), Zhao-Hua Lu (St. Jude Children's Research Hospital)*

Pediatric cancer treatment can have profound and complicated late effects. This is especially true for pediatric brain tumors. With the survival rates increasing as a result of improved detection and treatment, a more comprehensive understanding of the impact of current treatments on neurocognitive function and brain structure is critically needed. A frontline medulloblastoma clinical trial (SJMB03) has collected data, including treatment, clinical, neuroimaging, and cognitive variables. Advanced methods for modeling and integrating these data are critically needed to understand the disease etiology, treatment response, and long-term outcomes. We propose an integrative Bayesian mediation analysis approach to model jointly a treatment exposure, a high-dimensional structural neuroimaging mediator, and a neurocognitive outcome and to uncover the mediation pathway. For the SJMB03 study, the BI-GMRF method has identified white matter microstructure that is damaged by cancer-directed treatment and impacts late neurocognitive outcomes. The results provide guidance on improving treatment planning to minimize long-term cognitive sequela for pediatric brain tumor patients.

Random Forests for Time Series Forecasting and Forecast Intervals
Barbara Bailey (San Diego State University)*

Random forests have successfully been used for prediction in wide range of applications. Random forests consist of an ensemble of decision trees for regression or classification. The modeling and forecasting of time series data are investigated. The stationary bootstrap is implemented to generate realizations of the time series to be used in the building of each tree in the random forest and in the construction of forecast intervals.

Calibration Coefficient Estimation in Quantitative Fatty Acid Signature Analysis
Jennifer Mcnichol (University of New Brunswick)*

Quantitative fatty acid signature analysis (QFASA) has become a popular method of diet composition estimation for marine predators. Along with fatty acid signatures for a particular predator and their respective prey, QFASA requires calibration coefficients, which account for the differential metabolism of individual fatty acids. In practice calibration coefficients are not known and therefore must be estimated via feeding trails with captive animals. The main criticism of QFASA is that verifying the accuracy of calibration coefficients is nearly impossible and may introduce bias into the diet estimates. To resolve this issue, a new model was proposed which allows for estimation of calibration coefficients simultaneously alongside the diets. However, the proposed model has only a limited range of supporting evidence and has not yet been directly compared to traditional QFASA. In this talk, the results of a simulation study comparing the two approaches is presented.

SurvBenchmark: comprehensive benchmarking study of survival analysis methods using both omics data and clinical data
Yunwei Zhang (School of Mathematics and Statistics, the University of Sydney; Charles Perkins Centre, the University of Sydney), Jean Yang (School of Mathematics and Statistics, the University of Sydney; Charles Perkins Centre, the University of Sydney), Samuel Muller (School of Mathematics and Statistics, the University of Sydney;Department of Mathematics and Statistics, Macquarie University, Sydney)*

Survival analysis is a branch of statistics that deals simultaneously with both, the tracking of time and of the survival status as dependent response. Current comparisons of the performance of survival models mostly focus on classical clinical data with traditional statistical survival models, with prediction accuracy being often the only measurement of model performance. Moreover, survival analysis approaches for censored omics data has not been fully studied. To this end, we develop SurvBenchmark—a benchmark framework that enables comparison of both classical and state-of-art machine learning survival models for clinical and omics datasets using hybrid model evaluation metrics based on model predictability, stability, flexibility and computational efficiency.

Our results on 16 diverse collections of real-life data show that the model performance is data dependent and that there is no single method that performs the best across all assessment metrics. Furthermore, our results demonstrate varying performance for a given method between classical cox-based approaches and modern machine learning survival methods with potential to guide survival method selection.

Accurate Source Tracking Using Microbial Samples with Applications in Forensic Study
Qianwen Luo (The University of Arizona), Meng Lu (The University of Arizona), Kyle Carter (The University of Arizona), Hongmei Jiang (Northwestern University), Lingling An (The University of Arizona)*

Human microbiome has become popular in forensic studies due to its use in estimating the time since death, tracing evidence, and human identification. 16S rRNA sequencing technology or shotgun sequencing are used to detect and quantify bacteria in microbial samples. The wide use of human microbiome in forensic studies calls for advanced statistical and computational methods for analysis. We proposed a 4-step statistical procedure to link the microbial evidence collected in the crime scene to a group of suspects, then find the possible missing suspect, and finally estimate the proportion/contribution of each suspect. Through a series of comprehensive simulation studies, we have demonstrated that the new method outperformed the existing methods in trace evidence with smaller error rate.