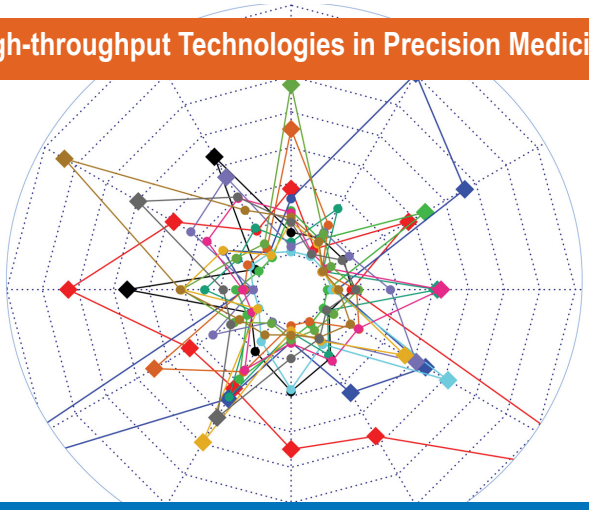
iBRIGHT 2019 Conference

Integrative Biostatistics Research for Imaging, Genomics, & High-throughput Technologies in Precision Medicine

Nov. 11-13 | Houston Texas

Cancer Biostatistics, Clinical Trials, Statistical Genetics, and Integrative Biostatistics Methods for High-throughput Omic Technologies (including Microbiomics)



ABSTRACTS

Plenary Presentations

Polygenic Modeling of Complex Traits to Inform Biology, Causality and Prediction

Nilanjan Chatterjee, Ph.D.
Johns Hopkins University

Using results from modern genome-wide association studies (GWAS), we and others have now unequivocally demonstrated that complex traits are extremely polygenic, with each individual trait potentially involving thousands to tens of thousands of genetic variants. While each individual variant may have a small effect on a given trait, in combination, they can explain substantial variation of the trait in the underlying population. In the past, analyses of GWAS have mainly focused on modelling genetic susceptibility one-variant-at-a-time, and identifying those which reach stringent statistical significance for association. In the future, however, we advocate that analysis needs to focus more on polygenic modelling to exploit the power of diffused signals in GWAS. In this talk, I will review recent advances in statistical methods for polygenic analysis, as well as scientific knowledge gained through their applications, in three areas of major interest (i) understanding biology through genomic enrichment analysis, (ii) exploring causality through Mendelian Randomization (Genetic Instrumental Variable) analysis, and (iii) and informing precision medicine through development of risk prediction models.

email: nchatte2@jhu.edu

Shotgun Metagenomics, Functional Microbiome and Its Applications in Cancer and Immunology

Hongzhe Li, Ph.D.
University of Pennsylvania

Shotgun metagenomic sequencing provides data for characterizing specific microbial strains and for understanding the functions of microbial communities. Functional microbiome aims to understand how microbial communities contribute to metabolic variability and host immunity by integrating other data types, such as metabolomics and flow cytometry data. I will present several statistical and computational methods for functional microbiome studies, including methods for identifying biosynthetic gene clusters and modeling of metabolic pathways, and methods for associating microbiome data with flow cytometry data. I will illustrate these methods using data from Crohn's disease and human cancer studies.

email: hongzhe@pennmedicine.upenn.edu

Augmenting Clinical Intelligence with Machine Intelligence

Suchi Saria, Ph.D.
Johns Hopkins University

email: ssaria@cs.jhu.edu

Genome-wide Models for Heritability and Prediction

David Balding, Ph.D.
University of Melbourne

Recently there has been much interest and great progress in statistical modelling of genome-wide GWAS test statistics for heritability analyses, genetic correlation estimates and to predict complex phenotypes, correcting for GWAS confounding bias, if required. The LDSC model has been widely applied since published in 2015. However, its implicit assumption of uniform expected heritability across SNPs, shared with the earlier GCTA model that

analyzed individual genotype data, is now recognized to be unrealistic and to have led to poor performance. The LDSC model has been further developed (now called S-LDSC), including SNP-specific expected heritabilities that adjust for effects of minor allele fraction, linkage disequilibrium, and genome annotations. We have proposed a different approach, implemented in the SumHer software, finding in some settings dramatically different inferences from those based on LDSC. As the models have improved, so have methods for model comparison, including an improved log likelihood approximation and an approach based on leave-one-chromosome-out prediction of summary statistics. We report latest developments, showing that as the models have improved, inferences are converging. However, we also report that the adjustment for GWAS confounding bias is unreliable in all current approaches, even if the assumed heritability model is correct, so that summary-statistic analyses should be limited to settings in which the original GWAS adequately adjusted for confounding.

email: david.balding@unimelb.edu.au

Session Presentations

1. Network Inference for Complex Biological Data

1a. Modeling Association in Microbial Communities with Clique Loglinear Models

Adrian Dobra, Ph.D.
University of Washington

There is a growing awareness of the important roles that microbial communities play in complex biological processes. Modern investigation of these often uses next generation sequencing of metagenomic samples to determine community composition. We propose a statistical technique based on clique log-linear models and Bayes model averaging to identify microbial components in a metagenomic sample at various taxonomic levels that have significant associations. We describe the model class, a stochastic search technique for model selection, and the calculation of estimates of posterior probabilities of interest. We demonstrate our approach using data from the Human Microbiome Project and from a study of the skin microbiome in chronic wound healing. Our technique also identifies significant dependencies among microbial components as evidence of possible microbial syntrophy.

Email: adobra@uw.edu

1b. Spike-and-Slab Priors for Multiple Graphs Estimation

Marina Vannucci, Ph.D.
Rice University

There is now a huge amount of literature on Bayesian methods for variable selection that use spike-and-slab priors. Such methods, in particular, have been quite successful for applications in a variety of different fields. A parallel development has happened in graphical models, where priors are specified on precision matrices, with recent developments in the estimation of multiple graphs that may share common features. In this talk I describe various prior constructions that account for the presence/absence of edges, or for the strength of the connections. I also address extensions of the models to the analysis of non-Gaussian data. I motivate the development of the models using specific applications from neuroimaging and from studies that use microbiome data.

Email: marina@rice.edu

1c. Estimating Dynamic Brain Functional Networks Using Multi-subject fMRI Data

Suprateek Kunda, Ph.D.
Emory University

A common assumption in the study of brain functional connectivity is that the brain network is stationary. However, it is increasingly recognized that the brain organization is prone to variations across the scanning session, fueling the need for dynamic connectivity approaches. One of the main challenges in developing such approaches is that the frequency and change points for the brain organization are unknown, with these changes potentially occurring frequently during the scanning session. In order to provide greater power to detect rapid connectivity changes, we propose a fully automated, two-stage approach that pools information across multiple subjects to estimate change points in functional connectivity, and subsequently estimates the brain networks within each state phase lying between consecutive change points. The number and positioning of the change points are unknown and learned from the data in the first stage, by modeling a time-dependent connectivity metric under a fused lasso approach. In the second stage, the brain functional network for each state phase is inferred via sparse inverse covariance matrices. We compare the performance of the method with existing dynamic connectivity approaches via extensive simulation studies, and we apply the proposed approach to a saccade block task fMRI data. Our approach is seen to be scalable to high dimensional brain networks and demonstrates the ability to accurately estimate the change points for the dynamic network.

Email: suprateek.kundu@emory.edu

---

2. Single Cell Technology

2a. Breast Cancer Evolution – Insights from Single Cell Genomics

Nick Navin, Ph.D.
The University of Texas MD Anderson Cancer Center

Email: NNavin@mdanderson.org

2b. Gene Expression Imputation and Clustering with Batch Effect Removal in Single-cell RNA-seq Analysis by Deep Learning

Mingyao Li, Ph.D.
University of Pennsylvania

A primary challenge in single-cell RNA-seq (scRNA-seq) analysis is the ever increasing number of cells, which can be thousands to millions of cells in large projects, such as the Human Cell Atlas. Identifying cell populations becomes challenging in these data, as many existing scRNA-seq clustering methods cannot be scaled up to handle such large datasets. For large data, it is desirable to learn cluster-specific gene expression signatures from the data itself. Another challenge in large-scale scRNA-seq analysis is batch effect, which refers to systematic gene expression difference from one batch to another. Failure to remove batch effect can obscure downstream analysis and interpretation of results. In this talk, I will present a method for scRNA-seq analysis that enables gene expression imputation and clustering simultaneously through the use a deep learning algorithm. We further extend this method to incorporate known cell type information from a well-labeled source dataset through the use of transfer learning, a machine learning method that transfers knowledge gained from one problem to a different, but related problem. Through comprehensive evaluations across many datasets generated in different tissues, species, and protocols, we show that our method can significantly improve clustering accuracy as compared to existing methods, and it is capable of removing complex batch effects, while maintaining true biological variations. We expect that, with the increasing growth of single-cell studies, our methods offer a useful set of tools for clustering of these data.

Email: mingyao@pennmedicine.upenn.edu

2c. Efficient Clustering and Differential Expression of Large Single-cell Datasets

Davide Risso, Ph.D.
University of Padova

Single-cell RNA-Seq (scRNA-seq) enables the genome-wide measurement of gene expression at the single-cell level. In many applications, the first step of the analysis is the identification of cell subpopulations, through the use of unsupervised clustering algorithms, such as k-means. However, current unsupervised clustering algorithms and corresponding software implementations can be slow and typically require the data to be loaded entirely into memory. This can be challenging with large compendium datasets, such as the Human Cell Atlas, the size of which ranges from thousands to millions of cells within a dataset. To address this, we developed an open-source implementation of the mini-batch k-means algorithm in the R/Bioconductor package mbkmeans. This implementation leverages new and existing Bioconductor facilities to work with on-disk data representations and delayed operations. Our mini-batch k-means clustering can be applied to several matrix-like data containers, including base R matrices, sparse matrices, delayed and HDF5 matrices. We will demonstrate the performance of the mbkmeans package on real and simulated large datasets. We will also highlight and compare the performance against regular k-means, density-based algorithms (such as shared nearest neighbor clustering), and other mini-batch k-means implementations. *(Joint work with: Stephanie Hicks, Elizabeth Purdom, Yuwei Ni, and Ruoxi Liu.)*

Email: dar2062@med.cornell.edu

---

3. Triumphs and Challenges in the Design and Conduct of Novel Adaptive Oncology Trials

3a. The STAMPEDE trial and beyond: improving outcomes as rapidly as possible for patients

Max Parmar, Ph.D.
University College, London

STAMPEDE is multi-arm, multi-stage platform randomized protocol for men with prostate cancer who are starting hormone therapy. Since its launch in 2005, STAMPEDE has recruited more than 11,000 men and produced 3 practice-changing results in 2015, 2017, and 2018. The control and research arms of STAMPEDE have been adapted to incorporate the changes in standard of care as a consequence of these results. STAMPEDE has led the way in the idea of dropping and adding research arms (in a phase III protocol), such that over a 20 year period it will evaluate 11 different new treatments. We describe the design, results, and current status of STAMPEDE. A key feature of STAMPEDE has been the ability to test new research arms in subsets of patients. We argue that this is the best way in which to evaluate new treatments (particularly within subsets of patients) and show how STAMPEDE and our other platform trials have developed to do this.

Email: m.parmar@ucl.ac.uk

3b. The Lung-MAP Master Protocol: Lessons Learned

Mary Redman, Ph.D.
Fred Hutchinson Cancer Research Center

The Lung-MAP master protocol is a study designed to evaluate biomarker-driven therapies and immunotherapies in previously-treated non-small cell lung cancer patients. Activated to accrual in June 2014, it was the first of the master protocols initiated in the National Clinical Trials Network of the National Cancer Institute. It was implemented using a public-private partnership and operation-

ally led by the SWOG Cancer Research Network. The protocol included a common biomarker screening component, which led eligible patients to enroll onto a biomarker-driven sub-study evaluating a therapy targeted against the biomarker or to enroll on to a "non-match" sub-study for patients without any of the matching biomarkers. Since activation, the study has enrolled over 2,000 patients to the screening studies. Within the protocol, almost 700 patients have been enrolled onto a treatment sub-study; seven sub-studies have been completed, three are currently open to accrual or waiting for data to mature for interim analysis, and at least four sub-studies are in active development, expected to activate within the year.

In this talk I will discuss our lessons learned from conducting a complicated study amidst a changing landscape. Implementation of a complicated protocol, such as Lung-MAP, requires constant effort and communication between all stakeholders, and a focused set of eyes on all aspects of the study at all times. In addition, there are practical challenges to maintaining the protocol with a continual flow of new study ideas and study development. I will provide some specific details on the steps we have taken to maintain the study infrastructure, while also being flexible to changes and the demands of clinical research.

Email: mredman@fredhutch.org

3c. Opportunities and Challenges in the Design and Implementation of Biomarker-driven Trials

Susan Halabi, Ph.D.
Duke University

Increased advances in understanding the roles of molecular and genetic pathways in carcinogenesis are leading to the development of novel therapies that target these pathways. Before embarking on a trial of a targeted agent that either prospectively stratifies or selects patients (i.e., enrichment) based on the relevant molecular marker, a strong scientific rationale is necessary. In this talk, issues in the design and implementation of biomolecular marker-based strategies in phase II and phase III trials in oncology will be discussed. In addition to describing the statistical design selected to evaluate efficacy of the targeted agent, the talk will also highlight the scientific rationale and compelling clinical data, including validation of the target pathways at the time the trials were designed. Statistical and logistical issues relating to the conduct and monitoring of such trials will also be discussed, and real life examples will be provided.

Email: susan.halabi@duke.edu

3d. Lessons Learned from Bayesian Adaptive Trials – The I-SPY2 Trial and Beyond

Scott Berry, Ph.D.
Berry Consultants

In this talk I will describe lessons learned from the I-SPY2 Bayesian adaptive platform trial and how these lessons are being implemented in other oncology adaptive platform trials, including adaptation into many other therapeutic areas. The GBM-AGILE (glioblastoma) and Precision Promise (pancreatic cancer) trials are phase III adaptive platform trials using many of the innovative features of I-SPY2 to bring much needed efficiency to these challenging therapeutic areas. In addition, adaptive platform trials are being conducted in areas such as Alzheimer's, ALS, infectious disease, and pediatric rare diseases, like Duchenne muscular dystrophy. New challenges to adaptive platform trial designs will be discussed, as well as solutions implemented in the trials.

Email: scott@berryconsultants.net

---

4. Pharmacogenomics: Current Challenges and Opportunities

4a. Precision Medicine in Oncology: Neither Hype nor Silver Bullet

Amber Johnson, Ph.D.
The University of Texas MD Anderson Cancer Center

Precision medicine can be defined as the use a biomarker to identify therapies for which an individual patient might receive therapeutic benefit. However, there are only ~50 biomarkers associated with an FDA-indication for an approved therapy to guide clinical decision making. Thus, we have developed a dynamic decision support system, Precision Oncology Decision Support (PODS), that allows us to provide clinicians with real-time actionability classification of biomarkers in their patients, regardless of the source of the testing. This system couples expert manual curation with natural language processing-assisted curation to yield functional annotations on alterations found in patient tumor samples. In coordination with proactive clinical trial alerts, and engaged follow-up with clinicians, we have seen a doubling of matching patients to clinical trials based on biomarker status, as well as an improvement in overall outcome. Since being established in 2015, we have annotated 18,492 variants and issued 6,942 reports. It is important to note that only 25% of variants annotated to date are considered therapeutically potentially actionable or actionable. Many patients are unable to be matched, not because of a lack of associated trial, but because of progression of disease, suggesting that earlier biomarker testing could improve clinical trial enrollment rates. Moreover, less than half of our patients have a biomarker that could be characterized as potentially actionable for any of the 2,784 drugs annotated in PODS, signifying the need for both innovative drug development strategies, as well as broader biomarker screening methods beyond traditional methods like NGS and immunohistochemistry.

Email: AMJohnson2@mdanderson.org

4b. Integrative Data Science Methods to Elucidate Pharmacologic Response Mechanisms

Gerald Higgins, Ph.D.
University of Michigan

New insights into the architecture and dynamics of the noncoding regulatory genome have transformed our understanding of the cornerstones of classical pharmacogenomics—pharmacokinetics (PK) and pharmacodynamics (PD). The newly emerging concept of the "pharmacoepigenome," broadly enabled by the 4D Nucleome Concept, involves regulators including enhancers, promoters and RNAs of gene expression (termed the "Regulome"), located in the noncoding genome. The Regulome is characterized by a hierarchy of stereotypic transcriptional domains, in which variation profoundly impacts drug response in humans. Transcriptional control consists of canonical 3D structures that include topologically associated domains (TADs). Different TADs are activated or suppressed by drugs in a cell-type specific manner. Drug-disease networks have been found to be tightly coupled so that gene variants significantly associated with a disease are identical to, or are found within, the same regulatory networks that determine medication-based therapeutic outcome. Thus, mutations that disrupt the spatial hierarchy of transcription within euchromatin not only convey disease risk, but also concomitant variability in drug and pharmacogenomic response. Pathways containing disease risk, drug response, and concomitant adverse event variants have the potential to better inform therapeutic options for patients, based on the emerging pharmacological basis of drug response and adverse drug events. Examples of current and future applications of Machine Learning in pharmacogenomics, include 1) identification of novel regulatory variants located in noncoding domains of the genome and their function as applied to pharmacoepigenomics; 2) patient pharmacological and adverse event/drug reaction response stratification from medical records; and 3) the mechanistic prediction of drug response, targets, and their interactions. In addition, we anticipate that in the future, deep learning will be widely used to predict personalized drug response and optimize medication selection and dosing, using knowledge extracted from large and complex molecular, clinical, epidemiological, and demographic datasets. *(Joint work with: Michael Savageau)*

Email: gehiggin@med.umich.edu

4c. Discovery of Anticancer Drug Combinations through Quantitative Pharmacology

Anil Korkut, Ph.D.
The University of Texas MD Anderson Cancer Center

Our research interests span a wide range of major problems in cancer biology, including resistance to targeted agents and discovery of rational combination therapies. We integrate perturbation biology experiments, genomics, network modeling, and multiplex imaging for anticancer drug combination discovery and mechanistic studies. We also work on the development of network inference and machine learning algorithms to link cancer omics data and phenotypic responses to therapy. I am going to present examples of our algorithm development efforts and systems biology applications for effective drug combination discovery.

Email: AKorkut@mdanderson.org

4d. A Bayesian Precision Medicine Framework for Calibrating Individualized Therapeutic Indices in Cancer

Veera Baladandayuthapani, Ph.D.
University of Michigan

The development and clinical implementation of evidence-based precision medicine strategies has become a realistic possibility, primarily due to the rapid accumulation of large-scale genomics and pharmacological data from diverse model systems: patients, cell lines, and drug perturbation studies. We introduce a novel Bayesian modeling framework called the individualized theRapeutic index (iRx) model to integrate high-throughput pharmacogenomic data across model systems. Our iRx model achieves three main goals: first, it exploits the conserved biology between patients and cell lines to calibrate therapeutic response of drugs in patients; second, it finds optimal cell-line avatars as proxies for patient(s); and finally, it identifies key genomic drivers explaining cell line-patient similarities. This is achieved through a semi-supervised learning approach, which conflates (unsupervised) sparse latent factor models with (supervised) penalized regression techniques. We illustrate and validate our approach using two existing clinical trial datasets in multiple myeloma and breast cancer studies. We show that our iRx model improves prediction accuracy compared to naive alternative approaches, and it consistently outperforms existing methods in literature in both real clinical examples, as well as in multiple simulation scenarios.

Email: veerab@umich.edu

---

5. Clinical Trial Designs Using Biomarkers, Pharmacokinetics and/or Pharmacodynamics

5a. Optimizing Confirmatory Clinical Trials with Master Protocols

Thomas Jaki, Ph.D.
Lancaster University

We design two-stage adaptive confirmatory clinical trials that use adaptation to find the subgroup of patients who will benefit from a treatment. We make use of a Bayesian decision theoretic framework to optimize the design parameters. Given a prespecified utility function that

takes the prevalence of the subpopulations into account, our proposal allows altering allocation of patients to arm in an interim analysis, ensuring efficient use of available resources to maximize the expected utility. This design includes adaptive enrichment and single-stage designs as special cases. We consider testing the elementary null hypotheses of disjoint subgroups and guarantee strong control of the family-wise error rate using the conditional error rate approach. We show results for traditional trials with multiplicity control, as well as the setting of umbrella trials in which no overall error control is desired. We present the results of simulation studies in a variety of cases to study the optimization of the design parameters and the effectiveness of the proposed design.

Email: jaki.thomas@gmail.com

5b. Bayesian Population Finding with Biomarkers in a Randomized Clinical Trial

Satoshi Morita, Ph.D.
Kyoto University

The identification of good predictive biomarkers allows investigators to optimize the target population for a new treatment. We propose a novel utility-based Bayesian population finding (BaPoFi) method to analyze data from a randomized clinical trial with the aim of finding a sensitive patient population. Our approach is based on casting the population finding process as a formal decision problem, together with a flexible probability model, Bayesian additive regression trees (BART), to summarize observed data. The proposed method evaluates enhanced treatment effects in patient subpopulations, based on counterfactual modeling of responses to a new treatment and control for each patient. In extensive simulation studies, we examine the operating characteristics of the proposed method. We compare with a Bayesian regression-based method that implements shrinkage estimates of subgroup-specific treatment effects. For illustration, we apply the proposed method to data from a randomized clinical trial.

Email: smorita@kuhp.kyoto-u.ac.jp

5c. INSIGhT: A Bayesian Adaptive Platform Trial to Develop Precision Medicines for Patients With Glioblastoma

Lorenzo Trippa, Ph.D.
Harvard SPH

Adequately prioritizing the numerous therapies and biomarkers available in late-stage testing for patients with glioblastoma (GBM) requires an efficient clinical testing platform. We developed and implemented INSIGhT (Individualized Screening Trial of Innovative Glioblastoma Therapy) as a novel adaptive platform trial (APT) to develop precision medicine approaches in GBM. INSIGhT compares experimental arms with a common control of standard concurrent temozolomide and radiation therapy, followed by adjuvant temozolomide. The primary end point is overall survival. Patients with newly diagnosed unmethylated GBM who are IDH R132H mutation negative and with genomic data available for biomarker grouping are eligible. At the initiation of INSIGhT, three experimental arms (neratinib, abemaciclib, and CC-115), each with a proposed genomic biomarker, are tested simultaneously. Initial randomization is equal across arms. As the trial progresses, randomization probabilities adapt on the basis of accumulating results using Bayesian estimation of the biomarker-specific probability of treatment impact on progression-free survival. Treatment arms may drop because of low probability of treatment impact on overall survival, and new arms may be added. Detailed information on the statistical model and randomization algorithm is provided to stimulate discussion on trial design choices more generally and provide an example for other investigators developing APTs.

Email: ltrippa@jimmy.harvard.edu

5d. A Nonparametric Bayesian Basket Trial Design

Peter Mueller, Ph.D.
The University of Texas at Austin

Targeted therapies on the basis of genomic aberrations analysis of the tumor have shown promising results in cancer prognosis and treatment. Regardless of tumor type, trials that match patients to targeted therapies for their particular genomic aberrations have become a mainstream direction of therapeutic management of patients with cancer. Therefore, finding the subpopulation of patients who can most benefit from an aberration-specific targeted therapy across multiple cancer types is important. We propose an adaptive Bayesian clinical trial design for patient allocation and subpopulation identification. We start with a decision theoretic approach, including a utility function and a probability model across all possible subpopulation models. The main features of the proposed design and population finding methods are the use of a flexible non-parametric Bayesian survival regression, based on a random covariate-dependent partition of patients, and decisions based on a flexible utility function that reflects the requirement of the clinicians appropriately and realistically, and the adaptive allocation of patients to their superior treatments. Through extensive simulation studies, the new method is demonstrated to achieve desirable operating characteristics and compares favorably against the alternatives.

Email: pmueller@math.utexas.edu

---

6. Cancer Biostatistics and Health Services Research

6a. Translating New Biomarkers for Cancer Risk and Early Detection Technologies into Outcomes that Matter

Ruth Etzioni, Ph.D.
Fred Hutchinson Cancer Research Center

New cancer screening technologies offer the potential to broaden the scope of cancer early detection and enhance its life-saving potential. Studies of novel technologies focus on diagnostic performance and do not generally address long-term outcomes that ultimately drive cancer screening policies and decisions. In this presentation, I develop a modeling framework for projecting from diagnostic performance to outcomes that matter. The framework rests on two key elements: a model for disease natural history and a concept for how early detection impacts survival. I discuss data needs and identifiability considerations in natural history estimation and present several options for survival benefit models. The framework is applied to project the lives saved and overdiagnoses associated with the use of novel reflex biomarkers for prostate cancer screening. Studies of these markers have suggested that they provide improved diagnostic performance over the commonly used PSA test and offer a reduction in unnecessary biopsies performed. Our model results confirm empirical diagnostic properties and indicate that the reflex tests have the potential to reduce the risk of overdiagnosis, but also induce a non-trivial reduction in lives saved. This work is joint with Roman Gulati and the Fred Hutch CISNET prostate cancer modeling group.

Email: retzioni@fredhutch.org

6b. Bias Reduction Approaches for Cancer Outcomes Research Conducted Using Electronic Health Records

Rebecca Hubbard, Ph.D.
University of Pennsylvania

Electronic Health Records (EHR) offer the opportunity to study populations and outcomes it would be difficult or impossible to access using traditional observational study designs. These data sources are particularly valuable in the context of cancer outcomes research, because they provide access to large populations of patients receiving cancer care in community practice, providing a window into real-world care and outcomes. However, EHR data are not collected for research purposes and, as a result, the pattern of assessment and quality of outcome ascertainment varies across patients and clinical settings. We demonstrate bias induced in cancer outcomes studies, due to these characteristics of EHR data, using simulations and analyses of real-world data from an EHR-based study of breast cancer survivors. Reducing bias in EHR-based studies is challenging, due to the difficulty of obtaining gold-standard validation data for cancer outcomes and variation in the quality of outcome ascertainment across clinical settings, referred to as the “portability problem” for EHR-based phenotyping. We explore alternative approaches to bias reduction in settings where validation data are or are not available and assuming constant phenotyping accuracy or allowing for variation in phenotyping accuracy across clinical settings. Employing appropriate approaches for bias reduction is critical to obtaining valid results from EHR. The objective of this presentation is to raise awareness of the opportunities and pitfalls associated with EHR-based cancer outcomes research and introduce tools to mitigate some of the challenges.

Email: rhubb@pennmedicine.upenn.edu

6c. Using Social Network Analysis to Estimate the Effect of Network Position and Peer Effects with Application to the Intensity of End-of-life Care in Cancer

James O’Malley, Ph.D.
The Dartmouth Institute

In this talk I develop methodology for decomposing social network effects into local and external components with respect to organizational affiliation. Two types of models are considered: 1) a physician peer effects model of the intensity of care (health care spending) on cancer patients’ cost of care near the end of life; 2) a model of the effect of the structural importance of physician network positions on health outcomes of cardiovascular disease patients who undergo implantable cardiac defibrillator (ICD) therapy. In (1), both endogenous and exogenous peer effects are considered, while in (2), the modification of the network position effects by other physician characteristics is considered. The results for both applications illustrate the potential utility of social physician networks towards explaining key patterns of regional variation in healthcare outcomes and costs.

Email: James.OMalley@Dartmouth.edu

6d. Terminal Trend Models for Censored Quality-of-life Measures, Cost and Survival Data with Applications in Cancer Survivorship and Osteoporosis

Tor D. Tosteson, Sc.D.
Dartmouth College

In end-of-life studies, the primary outcomes are often health-related quality of life (HRQoL) measures and/or cost data. Randomized trials and prospective cohorts typically recruit patients with advanced stage disease and follow them until death or the end of the study. An important feature of such studies is that, by design, some patients, but not all, are likely to die during the course of the study. This affects the interpretation of the conventional analysis of palliative care trials and suggests the need for specialized methods of analysis. We have developed a “terminal decline model” for palliative care trials that, by jointly modeling the time until death and longitudinal measures, leads to flexible interpretation and efficient analysis of end of life data. (Li, Tosteson, Bakitas, STMED 2012; Bakitas, Tosteson, et. al. JCO 2015; Li, Frost, Tosteson et. al., STMED 2017)

Email: tor.tosteson@dartmouth.edu

7. Integrative Microbiome with Genomics

7a. Gut Bacteria and the Intestinal Barrier in Cancer Therapy

Robert Jenq, MD
The University of Texas MD Anderson Cancer Center

Febrile neutropenia is a common complication of cytotoxic cancer therapy. Considered a medical emergency, it requires immediate evaluation and prompt initiation of empiric broad-spectrum antibiotics. Intestinal commensal bacteria are important contributors to the pathophysiology of febrile neutropenia and among the most commonly isolated bacteria from bloodstream microbiological cultures. The vast majority of patients with febrile neutropenia, however, will have negative blood culture results. How the intestinal microbiome could contribute in these cases of febrile neutropenia is unclear. In the hematopoietic stem cell transplant patient population and in mouse models, we examine how intestinal bacteria contribute to systemic inflammation following cytotoxic cancer therapy, and identify oral nutrition, bacterial fermentation, and mucolytic bacteria as important elements in this pathway.

Email: RRJenq@mdanderson.org

7b. Bayesian Variable Selection for Microbiome Data

Christine Peterson, Ph.D.
The University of Texas MD Anderson Cancer Center

Our goal is to identify specific organisms in the microbiome that are associated to outcomes of interest. There are a number of challenges in selecting relevant features from microbiome data, however, including the fact that the relative abundances have a unit sum constraint and that the observed taxa are related within a phylogenetic tree. We propose a Bayesian variable selection framework, which both accounts for the compositional constraint and the relatedness among taxa. Specifically, we favor the joint selection of taxa that are close in genomic distance via an Ising prior on the variable selection indicators. We illustrate the proposed method with simulation studies that demonstrate improved performance over existing variable selection approaches, as well as an application examining the association between the gut microbiome and obesity.

Email: CBPeterson@mdanderson.org

7c. Integrative Analysis of Microbiome and Other Genomic Data Types

Michael Wu, Ph.D.
Fred Hutchinson Cancer Research Center

Joint analysis of microbiome and other genomic data types offers to simultaneously improve power to identify novel associations and elucidate the mechanisms underlying established relationships with outcomes. However, microbiome data are subject to high dimensionality, compositionality, sparsity, phylogenetic constraints, and complexity of relationships among taxa. Combined with the myriad of challenges specific to other omics data types, how to conduct integrative analysis continues to pose a grand challenge. To move toward joint analysis, we propose development of methods for identifying individual genomic features related to the microbiome community structure and for harnessing these associations to facilitate the analysis of individual data types. Specifically, using kernels to capture microbiome community structure, we develop approaches for rapidly screening genomic features affecting beta diversity, both marginally and conditionally on other genomic features. We further discuss how to exploit other data types to improve modeling of the effects of microbiome and/or genomics on clinical outcomes of interest.

Email: mcwu@fredhutch.org

7d. Statistical Methods for Analysis of Microbiome Data

Zhengzheng Tang, Ph.D.,
University of Wisconsin, Madison

Human microbiome studies using high-throughput DNA sequencing generate compositional data with the absolute abundances of microbes not recoverable from sequence data alone. In compositional data analysis, each sample consists of proportions of various organisms with a unit sum constraint. This simple feature can lead traditional statistical treatments, when naively applied, to produce errant results. In addition, microbiome sequence data are overdispersed and sparse with many zeros. These important features require further development of methods for analysis of compositional data. This talk presents several of the latest developments in this area. Real microbiome studies are used to illustrate these methods, and several open questions will be discussed.

Email: ztang2@wisc.edu

8. Practical Bayesian Precision Medicine: Design and Analysis

8a. Utility-based Design for Randomized Trials with Ordinal Outcomes and Prognostic Subgroups: Application to Pre-surgery Nutritional Prehabilitation in Esophageal Cancer Patients

Thomas Murray, Ph.D.
University of Minnesota

I will discuss a precision medicine design for randomized comparative trials with ordinal outcomes and prognostic

subgroups. The design accounts for patient heterogeneity by allowing possibly different comparative conclusions within subgroups. The comparative testing criterion is based on utilities for the levels of the ordinal outcome and a Bayesian probability model. I will compare and contrast designs based on models that include treatment-subgroup interactions, one proportional odds model and another non-proportional odds model with a hierarchical prior that shrinks toward the proportional odds model, and a third design that assumes homogeneity and ignores possible treatment-subgroup interactions. The three approaches have been applied to construct group sequential designs for a trial of nutritional prehabilitation versus standard of care for esophageal cancer patients undergoing chemoradiation and surgery, including both untreated patients and salvage patients whose disease has recurred following previous therapy. A simulation study will be presented that compares the three designs, including evaluation of within-subgroup type I and II error probabilities under a variety of scenarios, including different combinations of treatment-subgroup interactions.

Email: murra484@umn.edu

8b. Optimizing Natural Killer Cell Doses for Heterogeneous Cancer Patients Based on Multiple Event Times

Juhee Lee, Ph.D.
University of California, Santa Cruz

A sequentially adaptive Bayesian design is presented for a clinical trial of cord blood derived natural killer cells to treat severe hematologic malignancies. Given six prognostic subgroups defined by disease type and severity, the goal is to optimize cell dose in each subgroup. The trial has five co-primary outcomes, the times to severe toxicity, cytokine release syndrome, disease progression or response, and death. The design assumes a multivariate Weibull regression model, with marginals depending on dose, subgroup, and patient frailties that induce association among the event times. Utilities of all possible combinations of the nonfatal outcomes over the first 100 days following cell infusion are elicited, with posterior mean utility used as a criterion to optimize dose. For each subgroup, the design stops accrual to doses having an unacceptably high death rate, and at the end of the trial selects the optimal safe dose. A simulation study is presented to validate the design's safety, ability to identify optimal doses, and robustness, and to compare it to a simplified design that ignores patient heterogeneity.

Email: juheelee@soe.ucsc.edu

8c. Subgroup-specific Dose Optimization in Phase I Trials Based on Time to Toxicity with Adaptive Subgroup Combination

Andrew Chapple, Ph.D.
Louisiana State University

A Bayesian design is presented that does precision dose finding based on time to toxicity in a phase I clinical trial with two or more patient subgroups. The design, called Sub-TITE, makes sequentially adaptive subgroup-specific decisions, while possibly combining subgroups that have similar estimated dose-toxicity curves. Decisions are based on posterior quantities computed under a logistic regression model for the probability of toxicity within a fixed follow-up period, as a function of dose and subgroup. Similarly to the time-to-event continual reassessment method (TITE-CRM, Cheung and Chappell), the Sub-TITE design downweights each patient's likelihood contribution using a function of follow-up time. Spike-and-slab priors are assumed for subgroup parameters, with latent subgroup combination variables included in the logistic model to allow different subgroups to be combined for dose finding, if they are homogeneous. This framework can be used in trials where clinicians have identified patient subgroups, but are not certain whether they will have different dose-toxicity curves. A simulation study shows that, when the dose-toxicity curves differ between all subgroups, Sub-TITE has superior performance compared with applying the TITE-CRM while ignoring subgroups, and it has slightly better performance than applying the TITE-CRM separately within subgroups or using the two-group maximum likelihood approach of Salter et al that borrows strength among the two groups. When two or more subgroups are truly homogeneous, but differ from other subgroups, the Sub-TITE design is substantially superior to either ignoring subgroups, running separate trials within all subgroups, or the maximum likelihood approach of Salter et al.

Email: achapp@lsuhsc.edu

8d. Bayesian Nonparametric Survival Regression for Optimizing Precision Dosing of Intravenous Busulfan in Allogeneic Stem Cell Transplantation

Yanxun Xu, Ph.D.
Johns Hopkins University

Allogeneic stem cell transplantation (allo-SCT) is now part of standard of care for acute leukemia (AL). To reduce toxicity of the pre-transplant conditioning regimen, intravenous busulfan is usually used as a preparative regimen for AL patients undergoing allo-SCT. Systemic busulfan exposure, characterized by the area under the plasma concentration versus time curve (AUC), is strongly associated with clinical outcome. An AUC that is too high is associated with severe toxicities, while an AUC that is too low carries increased risks of disease recurrence and failure to engraft. Consequently, an optimal AUC interval needs to be determined for therapeutic use. To address the possibility that busulfan pharmacokinetics and pharmacodynamics vary significantly with patient characteristics, we propose a tailored approach to determine optimal covariate-specific AUC intervals. To estimate these personalized AUC intervals, we apply a flexible Bayesian

nonparametric regression model based on a dependent Dirichlet process and Gaussian process, DDP-GP. Our analyses of a dataset of 151 patients identified optimal therapeutic intervals for AUC that varied substantively with age and whether the patient was in complete remission or had active disease at transplant. Extensive simulations to evaluate the DDP-GP model in similar settings showed that its performance compares favorably to alternative methods. We provide an R package, DDPGPSurv, that implements the DDP-GP model for a broad range of survival regression analyses.
Email: yxu.stat@gmail.com

---

9. Novel Adaptive Designs and Statistical Methods to Accelerate Development of Immunology

9a. Statistical Considerations for Immunotherapy Trials in Cancer

Sumithra Mandrekar, Ph.D.
Mayo Clinic

The talk will begin with a description of the challenges with tumor measurement-based endpoints, particularly any RECIST-based endpoints. A large majority of this deals with missing data issues. The talk will highlight the many issues we identified when dealing with RECIST-based endpoints with traditional agents, most of which applies to the setting of immunotherapy. The options for endpoints in immunotherapy trials, and an example of a trial that illustrates cure fraction and how to model those data will be presented. Restricted mean survival time is another endpoint to be considered for immunotherapy trials, due to its interpretability. This is especially true in the case of non-proportional hazard treatment effects. The impact of various design parameters on the power of the test, based on restricted mean survival time, will be presented. Finally the immune RECIST criteria (iRECIST) to measure tumor response in immunotherapy trials will be briefly presented.

Email: Mandrekar.Sumithra@mayo.edu

9b. A Bayesian Phase I-II Design for Immunotherapy Trials

Suyu Liu, Ph.D.
The University of Texas MD Anderson Cancer Center

Immunotherapy is an innovative treatment approach that stimulates a patient's immune system to fight cancer. It demonstrates characteristics distinct from conventional chemotherapy and stands to revolutionize cancer treatment. We propose a Bayesian phase I/II dose-finding design that incorporates the unique features of immunotherapy by simultaneously considering three outcomes: immune response, toxicity, and efficacy. The objective is to identify the biologically optimal dose, defined as the dose with the highest desirability in the risk-benefit tradeoff. An Emax model is used to describe the marginal distribution of the immune response. Conditional on the immune response, we jointly model toxicity and efficacy using a latent variable approach. Using the accumulating data, we adaptively randomize patients to experimental doses, based on the continuously updated model estimates. A simulation study shows that our proposed design has good operating characteristics in terms of selecting the target dose and allocating patients to the target dose.

Email: syliu@mdanderson.org

9c. Sequential Interim Analysis for Immunotherapy Oncology Trials with Restricted Mean Survival Time Endpoint

Ying Lu, Ph.D.
Stanford University

A progress-free survival (PFS) time is often the primary endpoint in phase 3 oncology trials. The treatment effect of immune-therapies often fails the proportional hazards assumption. A comparison of the restricted mean survival times (RMSTs) has been proposed as an alternative to the hazards ratio to evaluate the PFS benefits. One of the challenges for using RMST is planning sequential clinical trials. Murray and Tsiatis (1999 Biometrics) have presented the statistical formulation for an RMST trial sequential design. In this talk, we will discuss some practical considerations, including the timing of interim analysis, confidence intervals of the resulting estimated difference in RMSTs, sample size consideration, and optimizations. Simulation examples will be presented to illustrate our results.

Email: ylu1@stanford.edu

9d. Design Concept for a Confirmatory Basket Trial

Robert Beckman, MD
Georgetown University

The discovery of numerous molecular subtypes of common cancers creates small "niche indications" within what were formerly large, histology-based categories. Enrolling an adequate sample size in confirmatory trials of corresponding targeted agents may be challenging. To the extent that these molecular subtypes may be common across histologies, they may be pooled in "basket" trials across histologies. Generally, these basket trials have been used for early exploratory development. In contrast, we present a qualitative concept for a confirmatory basket trial. Individual indications in the basket are evaluated based on a surrogate endpoint, which may lead to their accelerated approval. Indications are prescreened based on external data and on internal surrogate data before final pooling, and may be eliminated, if required, as a negative indication could dilute a positive result in other indications. Concepts for controlling the false positive rate, in spite of this pruning of indications, are discussed.

The pooled result is evaluated based on a definitive approval endpoint, for which a statistically significant result is required for the pooled data, and consistent trends are required for each individual indication remaining in the basket. We hope to facilitate the development of targeted agents for molecularly defined subtypes of cancer by providing recommendations for surmounting the challenges in confirmatory basket trials that will improve access to therapy for defined patient subpopulations (including accelerated approvals based on surrogate endpoints), lower associated development costs, and enhance the ability of health authorities to evaluate the risk and benefit of targeted agents designed for these indications.

Email: RAB302@georgetown.edu

---

10. Integrative High-throughput Omics

10a. Top-Down Integrative Genomics for Colon Cancer Precision Therapeutics

Jeffrey Morris, Ph.D.
University of Pennsylvania

In recent decades, the world of cancer research has become completely transformed by the development of hypersensitive technologies taking detailed biological measurements of quantities previously unmeasurable, including multi-platform genomics data containing complementary molecular information at the DNA, RNA, protein, and epigenetic levels. These data contain biological insights into molecular-based diseases like cancer, and the extraction of this knowledge from the data can lead to novel precision therapeutic strategies with the potential to change clinical practice. Successful knowledge extraction from these data depends on integrative methods that can combine information across these complementary modalities to paint a more complete picture of the underlying biology. I will discuss several integrative genomic methods in the context of colorectal cancer. We have established an interdisciplinary working group, as part of the MD Anderson Colorectal Cancer moonshot, which is using a top-down integrative learning approach to discover, validate, and translate new precision therapeutic concepts. These efforts heavily depend on our ability to use integrative genomics methods to deeply characterize the molecular characteristics of four recently discovered and validated subtypes of colorectal cancer (Guinney, et al. 2015 Nature Medicine, >1200 citations already) that are reshaping how the biomedical community defines and studies colorectal cancer. I will describe the big picture schema underlying our integrative learning approach, which can serve as a template for other settings, and I will give examples of some methods we are developing and using in this process.

Email: Jeffrey.Morris@pennmedicine.upenn.edu

10b. Precision Medicine via Robust Nonparametric Machine Learning and Outcome-guided Clustering with Multi-omics Data integration

George Tseng, Sc.D.
University of Pittsburgh

In this talk, two new statistical learning methodologies with multi-omics data integration will be covered. In the first project, the focus is to overcome difficulty of replicating classification models in independent studies using different omics experimental platforms. We propose to extend a top-scoring-pair technique to transform the original data into order-based 0-1 matrixes and hybrid with existing advanced machine learning methods (e.g., random forest or SVM) for a nonparametric classification model. We extend the method to integrate multi-omics and multi-cohort training data and demonstrate the improved validation prediction accuracy by simulation and large-scale real applications. In the second project, we present an outcome-guided clustering method to identify disease subtypes associated with outcome using single or multi-omics data. We show some theoretical insight of the proposed semi-supervised method compared to popular interaction-term based methods for personalized medicine. With simulation and real applications, we demonstrate performance of the method as more suitable for precision medicine (i.e., disease subtypes characterized by single or multi-omics data that are predictive to outcome).

Email: ctseng@pitt.edu

10c. Cell Type-specific Differential Expression and eQTLs

Wei Sun, Ph.D.
Fred Hutchinson Cancer Research Center

RNA-seq data are often collected from bulk tissue samples, most of which comprise a heterogeneous population of different cell types. Many studies have demonstrated that studying cell type-specific gene expression and cell type composition is crucial for many scientific and clinical questions, for example, classifying neuron subtypes; identifying genes and cell types related to specific diseases, such as the Zika virus infection; understanding cancer biology; and highlighting resistance mechanisms for cancer treatment. However, most statistical methods for RNA-seq data analysis assume that the RNA-seq data are collected from tissue samples with homogenous cell populations. We have developed statistical methods and software packages to close this gap in RNA-seq data analysis. Specifically, our methods perform cell type-specific differential expression testing or gene expression quantitative trait loci (eQTL) mapping using bulk RNA-seq data. For eQTL mapping, we borrow information from both total expression and allele-specific expression.

Email: wsun@fredhutch.org

10d. Integrative Factorization of Bidimensionally Linked Matrices

Eric Lock, Ph.D.
University of Minnesota

Advances in molecular “omics” technologies have motivated new methods for integrating multiple sources of high-content biomedical data. However, most methods to integrate multiple data matrices only consider data shared vertically (one cohort on multiple platforms) or horizontally (different cohorts on a single platform). This is limiting for data that take the form of bidimensionally linked matrices (e.g., multiple cohorts measured on multiple platforms), which are increasingly common in biomedical studies. We propose BIDIFAC (Bidimensional Integrative Factorization) for integrative dimension reduction and signal approximation of bidimensionally linked data matrices. Our method factorizes the data into (i) globally shared, (ii) row-shared, (iii) column-shared, and (iv) single-matrix structural components, facilitating the investigation of shared and unique patterns of variability. We use a penalized objective function that extends the nuclear norm penalty, and determine penalties via random matrix theory. We apply our method to integrate mRNA and miRNA expression data across cancerous tumor samples and normal samples. R code is available at https://github.com/lockEF/bidifac.

Email: elock@umn.edu

---

11. Integrative Genomics and Metabolomics

11a. Genetics, Metabolomics and Cardiovascular Disease

Bing Yu, Ph.D
UT Health

Scores of circulating metabolic signatures are identified to assist clinical management of cardiovascular disease (CVD). In contrast, little progress is made to characterize metabolic changes that presage development of CVD. Leveraging the rich metabolomics resources from prospective cohorts, we identified multiple metabolic signatures for CVDs. For example, a metabolite risk score (MRS) consisting of 19 metabolites was associated with a two-fold increased risk of coronary heart disease (CHD). After integrating with the CHD polygenic risk score (GRS), we observed a four-fold increased risk (top vs. bottom quartiles) of CHD. MRS and GRS jointly improved CHD 30-year risk prediction performance. We also found three metabolic pathways associated with an average of 15% risk difference of heart failure (HF), and genetic loci of those metabolic pathways directly contributed to HF risk. In addition, we explored metabolite quantitative trait loci (metQTL) at a genome-wide scale, including single nucleotide variants (SNV) and structural variants (SV), and identified hundreds of metQTLs among multi-ethnic populations. For example, an SNV at the upstream of SLC51A was associated with increased levels of a steroid sulfate - androsterone sulfate, and co-localized with SLC51A expression levels in transverse colon and terminal ileum. Interestingly, a large 64kb deletion (UGT2B17 upstream) was also associated with increased levels of androsterone sulfate. The biological effect of androsterone sulfate remains unclear, but we found evidence supporting its role in HF and the aging process. Taken together, by incorporating genomics and metabolomics at multiple granularity, we identified pathways involved in various CVD pathophysiological processes, shedding light on the disease etiology.

Email: Bing.Yu@uth.tmc.edu

11b. Post-GWAS Data Integration Identifies Novel Risk Factors for Alzheimer’s Disease

Qiongshi Lu, Ph.D.
University of Wisconsin, Madison

Despite the findings in genome-wide association studies (GWAS) for late-onset Alzheimer’s disease (LOAD), our understanding of its genetic architecture is far from complete. Transcriptome-wide association analysis that integrates GWAS data with large-scale transcriptomic databases is a powerful method to study the genetic architecture of complex traits. However, it is challenging to effectively use transcriptomic information given limited and unbalanced sample sizes in different tissues. Here we introduce and apply UTMOST, a principled framework to jointly impute gene expression across multiple tissues and perform cross-tissue gene-level association analysis using GWAS summary statistics. Compared with single-tissue methods, UTMOST achieved 39% improvement in expression imputation accuracy and generated effective imputation models for 120% more genes in each tissue. A total of 69 genes reached the Bonferroni-corrected significance level in the transcriptome-wide association meta-analysis for LOAD. Among these findings, we identified novel risk genes at known LOAD-associated loci, as well as five novel risk loci. Several genes, including IL10 and ADRA1A, also have therapeutic potential to improve neurodegeneration. Cross-tissue conditional analysis further fine-mapped IL10 as the functional gene at the CR1 locus, a well-replicated risk locus for LOAD. Extension of this framework to perform biobank-wide and metabolome-wide association studies will also be discussed. Overall, integrated analysis of omics annotations and biobank information provides insights into the genetic basis of LOAD and may guide functional studies in the future.

Email: qlu@biostat.wisc.edu

11c. Deep Learning and Spatial Modeling in Pathology Image Analysis

Guanghua Xiao, Ph.D.

UT Southwestern Medical Center

With the advance of technology, tumor tissue histology slide scanning is becoming a routine clinical procedure, which produces massive digital pathological images that capture histological details in high resolution. Reliable computational methods to predict patient prognosis and treatment response using tumor pathological slides will have an immediate impact on patient care in cancer. The spatial organization of different types of cells in tumor tissues reveals important information about the tumor microenvironment. In order to facilitate the study of cellular spatial organization and interactions, we developed computational algorithms to predict patient outcomes and analysis tools to characterize the tumor microenvironment from standard Hematoxylin and Eosin (H&E)-stained pathology images.

Email: Guanghua.Xiao@UTSouthwestern.edu

---

12. Statistical Genetics

12a. Machine Learning Methods with Applications to Genetic Analysis

Chris Amos, Ph.D.
Baylor College of Medicine

Machine learning tools have been effective in some cases in identifying patterns in data that are relevant for understanding disease pathogenesis. We completed a genome-wide association screening and performed a validation study to understand the genetic basis of a rare, but very severe, autoimmune disease called granulomatous polyangiitis. The genetic factors yielded an 80% attributable risk fraction, but included a large number of alleles in the HLA region, and we also wondered if the loci interacted to further increase risk for selected individuals. We applied Random Forest and Classification and Regression Trees, which identified some unanticipated interactions within and among loci. Results of this study allowed us to better understand combinations of variants that indicated greatly increased risks for some individuals. Although machine learning was effective, we then wondered if a more structured statistical approach would have yielded equivalent or more accurate results for characterizing risk. Classification trees provide clear algorithms for sub-setting individuals, but do not indicate if alleles that affect individuals at a population level also affect individuals within these subsets. We therefore built several new models to evaluate performance of more usual logistic regression approaches that modeled interactions among variables. Results showed that a logistic regression model of the classification tree provided slightly inferior prediction compared with the classification tree, but jointly modeling interactions using logistic regression led to a slightly better prediction accuracy. Current simulations are built from the most accurate classification and logistic regression findings to evaluate which approach, machine learning versus logistic regressions, provides a better prediction of risk when penalties for overfitting are considered in the model.

Email: Chris.Amos@bcm.edu

12b. Statistical Genomics of Clonal Mosaicism

Paul Scheet, Ph.D.
The University of Texas MD Anderson Cancer Center

Somatically-acquired allelic imbalance (AI) (i.e., chromosomal duplications, deletions, or copy-neutral loss-of-heterozygosity), is an established factor in cancer initiation and has recently been implicated as a marker for cancer risk. While DNA microarrays and next-generation sequencing are effective for whole-genome profiling of AI, in typical settings their sensitivities become extremely limited when the aberrant cell fraction (or tumor purity) is below 10-20%. Yet this range may be critical for early detection, diagnostics, or studies of pre-disease tissue, 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. Here, we present a haplotype-based statistical technique to powerfully detect these alterations and demonstrate its utility in the following challenging settings: tumors with high stromal content, premalignant lesions, normal “field” tissues proximal to a primary tumor, and blood.

Email: PAScheet@mdanderson.org

12c. Measuring Inter-individual Variation in Risk due to Genetic and Environmental Factors – Applications in Epidemiology of Nevi

Jaya Satagopan, Ph.D.
Rutgers School of Public Health

Nevi are among the important known risk factors for melanoma and increase with age during childhood and adolescence. Several studies have implicated genetic and environmental factors in the etiology of nevi and melanoma in adults. Inter-individual variation in disease or disease-related outcome is an important epidemiologic measure to weigh the impact of exposures or genetic factors on the outcome and to assess potential prevention programs. For example, if most of the variation is due to sun exposure, encouraging behavioral changes, such as reducing sun exposure or increasing sun protection, may need to be investigated as a potential secondary prevention strategy. However, if most of the variation is due to genetic factors, approaches for screening high-risk individuals based on genetic factors may need to be investigated. A natural measure of inter-individual variation in nevus counts is their statistical variance. However, reporting inter-individual variation in terms of statistical variance alone ignores the actual magnitude of the mean. This talk examines two alternative measures of variation:

the squared coefficient of variation and the Gini index. The properties of these measures, their relationship via a monotonic transformation, visual representations in terms of cumulative probability distribution functions of nevus counts, and relationship to epidemiologic measures of risk are described. A permutation test is presented to assess the statistical significance of additional variation explained by sun exposure vs. the variation explained by genetic factors. These concepts are applied to data on nevus counts, demographic factors, sun exposure, and multiple genetic factors from the Study of Nevi in Children by accounting for over-dispersion and by using penalized regression via the LASSO. Our results show that nevus development in early childhood is under strong genetic control and that additional risk factors for the development of nevi in pre-adolescent children remain to be identified.

Email: satagopj@sph.rutgers.edu

---

13. Electronic Health Records (EHR) Research

13a. Distributed learning from multiple EHR databases for predicting medical events

Qi Long, Ph.D.
University of Pennsylvania

Electronic health records (EHRs) data offer great promises in personalized medicine. However, EHRs data also present analytical challenges, due to their irregularity and complexity. In addition, analyzing EHR data involves privacy issues, and sharing such data across multiple institutions/sites may be infeasible. Building on a contextual embedding model, we propose a distributed learning method to learn from multiple EHRs databases and build predictive models for multiple diagnoses simultaneously, which can use both structured and unstructured data. We also augment the proposed method with Differential Privacy to further enhance data privacy protection. Our numerical studies demonstrate that the proposed method can build predictive models in a distributed fashion with privacy protection and the resulting models achieve comparable prediction accuracy compared with existing methods that use pooled data across all sites. This is joint work with Ziyi Li, Kirk Roberts, and Xiaoqian Jiang.

Email: qlong@pennmedicine.upenn.edu

13b. Individualized Treatment Recommendation (ITR) for Survival Outcomes

Haoda Fu, Ph.D.
Eli Lilly

ITR is a method to recommend treatment based on individual patient characteristics to maximize clinical benefit. During the past a few years, we have developed and published methods on this topic with various applications, including comprehensive search algorithms, tree methods, benefit risk algorithms, as well as multiple treatment and ordinal treatment algorithms. In this talk, we propose a new ITR method to handle survival outcomes for multiple treatments. This new model enjoys the following practical and theoretical features:

- Instead of fitting the data, our method directly searches the optimal treatment policy, which improves the efficiency.
- To adjust censoring, we have proposed a doubly robust estimator. Our method only requires that either the censoring model or survival model is correct, but not both. When both are correct, our method enjoys better efficiency.
- Our method handles multiple treatments with intuitive geometry explanations.
- Our method is Fisher's consistent, even under the misspecification of either the censoring model or the survival model (but not both).

Email: fu_haoda@lilly.com

13c. The AI Revolution is Coming to Medicine

Jiajie Zhang, Ph.D.
UT Health

Over the past 20 years, information technology has disrupted and transformed many major industries, including information retrieval, communication, retail, travel, finance, and education. With the recent breakthroughs in artificial intelligence, supported by explosive growth of data and computing power, information technology is finally beginning to solve some really hard problems in medicine. In this presentation, Dr. Zhang will discuss the opportunities and challenges in the application of AI in medicine.

Email: jiajie.zhang@uth.tmc.edu