Perelman School of Medicine
University of Pennsylvania

DBEI & CCEB RESEARCH DAY

A JOINT PROJECT OF

The Department of Biostatistics, Epidemiology and Informatics

AND

The Center for Clinical Epidemiology and Biostatistics

March 21, 2019 | 2nd ANNUAL EVENT
#2019ResearchDay

Biomedical Research Building (BRB)
Gaulton Auditorium
421 Curie Boulevard
Philadelphia, PA

ABSTRACTS

Welcome

Thank you for joining us at the second annual DBEI & CCEB Research Day! This event offers a snapshot of the latest research by the Department of Biostatistics, Epidemiology and Informatics and the Center for Clinical Epidemiology and Biostatistics and features this year's Brian L. Strom Visiting Professorship Lecture.

The DBEI distinctively brings together expertise in biostatistics, epidemiology and informatics, to advance its mission:
To discover, teach and promote impactful ways to preserve health, manage chronic disease and treat acute illness, by capitalizing on synergies across our three scientific disciplines.

The CCEB is an interdisciplinary and interdepartmental program that links clinical epidemiology and biostatistics within the Perelman School of Medicine, the University of Pennsylvania Health System, and the Penn community to advance its mission:
To foster research and training in clinical epidemiology and biostatistics, and serve as a resource to the clinical-research community.

1 Abstract

Spike and-Slab Group LASSOs for Grouped Regression and Sparse Generalized Additive Models

Ray Bai[1], Gemma Moran[1], Joseph Antonelli[2], Yong Chen[1], Mary R. Boland[1]

1. University of Pennsylvania
2. University of Florida

We introduce the spike-and-slab group LASSO (SSGL) for Bayesian estimation and variable selection in linear regression with grouped variables. We further extend the SSGL to sparse generalized additive models (GAMs), thereby allowing our model to flexibly capture nonlinear relationships. We develop highly efficient and scalable coordinate ascent algorithms for our model, and we illustrate our methodology through extensive simulations and data analysis on a Bardet-Biedl syndrome (BBS) case study. BBS is an autosomal recessive disorder which leads to progressive vision loss, and we aim to model the association between BBS and a set of genetic markers.

2 Abstract

Harmonization of Multi-Site Longitudinal MRI Neuroimaging Data

Joanne Beer[1], RT Shinohara[1], KA Linn[1]

1. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania

Aggregation of neuroimaging datasets from multiple sites is becoming increasingly common. While this presents opportunities for increased statistical power, it also presents challenges due to systematic scanner effects. We propose a method for the harmonization of multi-site longitudinal MRI data based on ComBat, a method originally developed for genomics and later adapted to cross-sectional MRI data. Using longitudinal cortical thickness data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), we demonstrate the presence of scanner-specific location and scale effects. We show that longitudinal ComBat can improve estimates of the association between baseline diagnosis group and change in cortical thickness over time in the ADNI data and via simulations.

3 Abstract

A Semi-Parametric Approach to Analyzing Error-Prone Failure Time Outcomes and Exposures

Lillian A. Boe[1], Pamela A. Shaw[1]

1. University of Pennsylvania

In clinical research, measurement error arises commonly in settings that rely on data from electronic health records or large observational cohorts. In particular, self-reported outcomes are typical in cohort studies for chronic diseases such as diabetes in order to avoid the burden of expensive diagnostic tests. Dietary intake, which is also commonly collected by self-report and subject to measurement error, is a major factor linked to diabetes and a number of other chronic diseases. These errors can bias exposure-disease associations that ultimately impact clinical decision-making. To address this bias problem, we have extended an existing semiparametric likelihood-based method for handling error-prone, discrete failure time outcomes to also address covariate measurement error. We conduct an extensive numerical study to evaluate the proposed method in terms of bias and mean-squared error in the estimation of the regression parameter of interest. This method is applied to data based from the Hispanic Community Health Study/Study of Latinos, a large population-based cohort study designed to examine risk factors for chronic diseases. Upon implementing the proposed method, we are able to assess the effect of dietary intake on the risk of incident diabetes mellitus when both variables are measured by self-report and hence subject to error.

4 Abstract

Mapping Regional Effects of Exposure to Hydraulic Fracturing Fluid and Linking with Information on Toxicity

Mary Regina Boland[1-4], Caroline DeVoto[1-4]

1. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania
2. Institute for Biomedical Informatics, University of Pennsylvania
3. Center for Excellence in Environmental Toxicology, University of Pennsylvania
4. Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia

Context/Purpose: Exposure to hydraulic fracturing fluid has been linked to multiple health conditions, including preterm birth. However, large-scale analyses of exposure to various toxic chemicals found within fracturing fluid remains under-explored.

Methods: We used data from FracFocus (https://fracfocus.org/) on hydraulic fracturing fluids and their known chemical ingredients and linked this information with the CDC's Agency for Toxic Substances and Disease Registry (ATSDR). We calculated the total number of active wells with chemicals in the ATSDR per state and per year. We also linked chemical ingredients with proteins affected using the Toxin Exposome Database (http://www.t3db.ca/).

Results: We mapped out the regions affected by exposure to specific chemicals, both those listed as toxic by the ATSDR versus those listed as non-toxic. We also mapped the effects of chemicals on various testosterone and estrogen pathways as recorded in the Toxin Exposome Database. We report regional differences.

Interpretation: Exposure to toxins in hydraulic fracturing fluid is important for exposome related research and also when comparing geographic regions to each other.

Conclusion: Exposure to hydraulic fracturing fluids may be important in understanding an individual's entire exposome. Some clustering by geographic region was observed for certain fracturing compounds.

5 Abstract*

GWAS of Hippocampal Subfield and Neighboring Cortical Structure Volumes Identifies an ERC1 Locus Using ADNI High-Resolution MRI Data

S Cong[1], X Yao[2], K Nho[3], SL Risacher[3], AJ Saykin[3], L Shen[2]

1. School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana
2. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania
3. Department of Radiology and Imaging Sciences, School of Medicine, Indiana University, Indianapolis

Given the heterogeneous structure of the human hippocampal complex and its relevance to Alzheimer's disease (AD), we performed a genome-wide association study (GWAS) of hippocampal subfield volumes as well as those of neighboring cortical structures using high-resolution MRI with the goal of determining genetic contributions to structural changes in this important region.

Participants included 136 ADNI subjects with both high-resolution MRI scans and genotype data available. Automatic Segmentation of Hippocampal Subfields (ASHS) software was employed to segment 14 primary labeled regions including hippocampal subfields and neighboring cortical structures. GWAS were performed to examine associations between 565,373 SNPs and 14 volumetric measures, with age, gender, education, ICV and diagnosis as covariates. We identified a novel locus rs2968869 in ERC1 on Chr. 12 ($p = 3.43E-9$; Bonferroni corrected $p = 2.71E-2$) significantly associated with right BA36 volume (Fig. 1). Right BA36 volume was previously reported that associated with tau deposition in right temporal lobe. This evidence indicates that BA36 might play a role in AD pathology. The minor allele C of rs2968869 was associated with greater right BA36 volume, suggesting a protective effect for AD. This aligns well with the IGAP finding that rs2968869 is a protective locus for AD ($p = 4.47E-2$).

High-resolution MRI provides detailed neuroanatomical information on hippocampal subfields structures. GWAS of these neuroimaging phenotypes identified a novel locus in ERC1 which is in regulation of neurotransmitter release and the NF-kappaB signaling pathway. After independent replication, ERC1 could be explored as a potential therapeutic target.

6 Abstract

Differences in Reported Symptom Type and Time to Recover Among Women and Men in the Ivy League-Big Ten Epidemiology of Concussion Study

B D'Alonzo[1], Carrie Esopenko[2], D Smith[3], D Wiebe[1]

1. Penn Injury Science Center and DBEI, University of Pennsylvania
2. Department of Rehabilitation and Movement Sciences, Rutgers University
3. Penn Center for Brain Injury and Repair, Department of Neurosurgery, University of Pennsylvania

Background: Research on gender differences in concussion and recovery is limited to small clinical studies. We examine gender differences in reported symptom type and time to recover within a large concussion surveillance system of varsity athletes.

Methods: Data on 22 symptoms and demographics were collected by athletic trainers who identified concussed athletes. Trainers monitored athletes for dates of symptom resolution through return to full play. We use factor analysis to identify symptom domains by gender and compare Kaplan-Meier survival curves for women and men.

Results: 1,243 athletes—512 women (39.8%), 731 men (60.2%)—representing 28 sports were included in the sample. Symptoms had excellent internal consistency in women (alpha=0.87) and men (alpha=0.86). Total symptoms did not differ (median count women=12, men=11). In women, 22 symptoms loaded on 6 domains representing emotional/anxious, fatigue, dizzy/blurred vision, headache/sensitivity, pressure/don't feel right/no nausea, and neck pain. In men, symptoms also loaded on 6 domains, but with different symptom groupings, where sleep co-occurred with nausea, and neck pain with pressure. Symptom resolution was slower ($p<0.05$) for women (median=14 days) than men (median=10 days) when considering cases with the most severe symptom profiles (5 or 6 of symptom domains types). After excluding football athletes, symptom resolution did not differ by sex (median days women=14 men=13).

Conclusions: Women and men report slightly different symptom experiences. Although women appear to take longer than men for symptoms to resolve, these differences disappear when football players are excluded. Further investigation into factors that may make football athletes different is needed.

7 Abstract

Integration of ChIP-Seq Data Identifies Global and Cell-Specific Glucocorticoid Receptor-Mediated Transcriptomic Changes

Avantika Diwadkar[1], Mengyuan Kan[1], Blanca E. Himes[1]

1. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania

ChIP-Seq, a technique that allows for in-depth quantification of DNA sequences bound by transcription factors or histones, has been widely used to characterize genome-wide DNA-protein binding induced by specific exposures and conditions. Over 40,000 ChIP-Seq studies of various DNA-binding proteins are available in public repositories. Integrating results of multiple ChIP-Seq datasets offers a cost-effect avenue to identify robust DNA-protein binding sites and determine their cell-type specificity. We developed brocade, a computational pipeline for reproducible analysis of publicly available ChIP-Seq data, that facilitates creation of R markdown reports with information on public datasets downloaded, quality control, and differential binding comparisons. We used brocade to analyze ChIP-Seq datasets of glucocorticoid receptor (GR), a transcription factor that mediates transcriptional response to glucocorticoids, commonly used anti-inflammatory drugs. Specifically, we analyzed ChIP-Seq studies of airway smooth muscle, airway epithelial cells, A549 cells, childhood acute lymphoblastic leukemia cells, and lymphoblastoid cells to identify cell type-specific and global GR binding across the five cell types. Our results demonstrate the utility of the brocade pipeline and identify GR binding sites that may mediate tissue-specific glucocorticoid responses.

8 Abstract

A Local Group Differences Test for Subject-Level Multivariate Density Neuroimaging Outcomes

JD Dworkin[1], KA Linn [1], TD Satterthwaite[2], A Raznahan[3], R Bakshi[4], RT Shinohara[1]

1. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
2. Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
3. Developmental Neurogenomics Unit, National Institute of Mental Health, National Institutes of Health, Bethesda, MD
4. Departments of Neurology and Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

Much of neuroimaging research focuses on voxel-wise analysis or segmentation of tissue damage, yet many diseases are characterized by diffuse or non-localized processes in brain tissue. In simple cases these processes can be quantified using summary statistics of voxel intensities. However, the manifestation of a disease process on imaging data is often unknown, or appears as a complex and non-linear relationship between the voxel intensities on various modalities. When the relevant pattern is unknown, the use of summary statistics is at best unable to capture differences between disease groups, and at worst encourages post-hoc searches for the optimal summary measure. In this study, we introduce a method for the naive discovery of group differences in voxel intensity profiles. The method operationalizes multi-modal magnetic resonance imaging data as multivariate subject-level densities of voxel intensities, and utilizes kernel density estimation to develop a local two-sample test for individual points within the density space. Through simulations, we show that this method controls type I error and recovers relevant differences when applied to a single point. Additionally, we demonstrate the ability to control family-wise error and maintain power when applying the test over increasingly fine grids within the density space. Finally, we apply this method to a study of subjects with either relapsing-remitting or secondary-progressive multiple sclerosis, and find significant differences between these disease subtypes in voxel intensity profiles within the thalamus.

Abstract

Regularized Prediction Modeling in Small Samples with Application to Predicting Toxicity in a CAR T-Cell Immunotherapy Trial

M Edmondson[1], D Teachey[2], P Shaw[1]

1. University of Pennsylvania
2. Children's Hospital of Philadelphia

The continual advent of novel therapies for treating cancer has increased the need for early phase clinical trials to determine the efficacy and safety of particular therapies for a given patient. In the age of personalized medicine, identification of biomarkers predictive of toxicity can improve patient management and potentially identify novel therapeutic targets. Risk prediction models are therefore often used to estimate risk of toxicity in a patient given levels of certain biomarkers. Analytical approaches for constructing these models, including various regularized logistic regression techniques, have most often been evaluated for large samples. In the setting of early phase clinical trials, however, samples are typically small, creating a need for evaluation of these methods in a small-sample setting. Through a structured simulation study, we compare the performance of several regularized logistic regression procedures in a high-dimensional, small-sample setting: stepwise regression, lasso, elastic net, and ridge regression. Our analytic approach can be easily adapted to other settings, allowing researchers to select a prediction modeling approach best-suited for a particular small-sample study. We apply these methods to select an approach to predict risk of cytokine release syndrome following treatment with CAR T-cell immunotherapy in a pediatric cohort with acute lymphocytic leukemia.

10 Abstract

Exploring Virtual Realty for Older Adults

C Fairman[1], S Oh[2], P Cacchione[2], G Demiris[2,3]

1. School of Engineering and Applied Science, University of Pennsylvania
2. School of Nursing, University of Pennsylvania
3. Perelman School of Medicine, University of Pennsylvania

Virtual Reality (VR) Applications have the potential to engage older adults who may be experiencing mild cognitive impairment and provide meaningful and engaging experiences especially for those facing social isolation or loneliness. While commercially available VR products are targeting the older adult population, little is known as to both anticipated benefits and challenges as well as older adults' needs and preferences pertaining to VR tools and their functionalities and attributes. This work presents a scoping review of available VR systems targeting older adults and the clinical domains they address, as well as preliminary work in the design, implementation and testing of a VR prototype designed specifically to address older adult users and provide features that facilitate engagement, entertainment and potentially reminiscence therapy. We describe a qualitative pilot study to assess older adult users' experience with the prototype and their preferences and needs. Furthermore, we provide a set of recommendations for user-centered design of VR systems in gerontology based on the specifics of the user groups and the clinical/ behavioral domains that such systems would address. We also highlight safety risks and ethical challenges associated with the use of VR informatics tools in gerontology.

11 Abstract*

Incorporating Single-Cell RNA-Seq Data to Infer Allele-Specific Expression

Jiaxin Fan[1], Rui Xiao[1], Mingyao Li[1]

1. University of Pennsylvania

Allele-specific expression (ASE) can be quantified by the relative expression of two alleles in a diploid individual, and such expression imbalance may explain phenotypic variation and disease pathophysiology. Existing methods detect ASE using easily obtainable bulk RNA-seq data, a data type that averages out possible heterogeneity in a mixture of different cell types. Since ASE may vary across different cell types, with the recent advance in single-cell RNA sequencing (scRNA-seq), characterizing ASE at the cell type resolution may help reveal more about the gene regulation. However, scRNA-seq data is costly to generate and noisy with excessive zeros due to transcriptional bursting. Therefore, it is desirable to incorporate information obtained from scRNA-seq data together with bulk data to infer cell type specific ASE. By employing cell type deconvolution and simultaneously modeling of multi-individual information, we are able to detect cell type specific ASE. Extensive simulations indicate that our method performs consistently well under a variety of scenarios.

12 Abstract

Design and Analysis of Two-Phase Samples in Discrete-Time Survival Analysis with Error-Prone Exposures

Kyunghee Han[1], Thomas Lumley[2], Bryan E. Shepherd[3], Pamela A. Shaw[1]

1. University of Pennsylvania
2. University of Auckland
3. Vanderbilt University

Increasingly medical research is dependent on data collected for non-research purposes, such as electronic health records data (EHR). EHR data and other large databases can be prone to measurement error in key exposures. Validating a subset of records is a cost-effective way of gaining information on the error structure, which in turn can be used to adjust analyses for this error and improve inference. We extend the mean score method for the two-stage analysis of discrete-time survival models, which uses the unvalidated covariates as auxiliary variables that can act as surrogates for the unobserved true exposure. This method allows for a two-phase sampling analysis approach that preserves the consistency of the regression model estimates in the validated subset, with increased precision leveraged from the auxiliary data. Further, we develop optimal sampling strategies which minimize the variance of the mean score estimator for a target exposure under a fixed cost constraint. Through simulations, we evaluate efficiency gains of the mean score estimator using optimal validation designs compared to random sampling. We also apply the proposed method to data from a large observational HIV cohort.

13 Abstract*

Effect of Anesthesia on Cardiac Hemodynamics in Patients Undergoing Durable Left Ventricular Assist Device Implantation: The EACH-LVAD Study

T Hanff[1], P Patel[2], K Kurcik[1], S Rao[1], S Kimmel[3], M Putt[3], P Atluri[4], C Bermudez[4], M Acker[4], E Birati[1], J Rame[1], J Wald[1]

1. Division of Cardiology, University of Pennsylvania
2. Department of Anesthesiology and Critical Care, University of Pennsylvania
3. Department of Biostatistics and Epidemiology, University of Pennsylvania
4. Division of Cardiovascular Surgery, University of Pennsylvania

Background: Right ventricular (RV) dysfunction is often unmasked or exacerbated during left ventricular assist device (LVAD) implantation. Prior to LVAD implantation, the effect of anesthesia induction on RV performance is unknown. We assessed for early hemodynamic changes in the RV as a function of anesthesia induction agent, prior to initiation of cardiopulmonary bypass, among patients receiving an LVAD.

Methods: We prospectively collected preoperative and intraoperative hemodynamics in patients undergoing LVAD implantation at our institution between 9/2017 and 9/2018. Preoperative RV hemodynamics were compared to RV hemodynamics post-induction and intubation but prior to bypass, including central venous pressure (CVP), mean pulmonary artery pressure (mPAP), and pulmonary artery pulsatility index (PAPi). The association of induction agents with serial changes in hemodynamics was assessed in a mixed effects model adjusting for use of pre-operative temporary mechanical circulatory support (tMCS) and vasoactive inotropic score.

Results: 41 patients were analyzed, with 4 intubated prior to induction. There was a significant average increase in CVP (13.9 mmHg, $p<0.001$) and mPAP (13.8 mmHg, $p<0.001$) and decrease in PAPi (-4.07, $p<0.001$) from induction to 60 minutes post-induction. Average PAPi was worse at all times among patients receiving preoperative tMCS (-1.8, $p<0.001$). No association was seen between choice or dose of anesthetic and hemodynamic changes.

Conclusion: RV hemodynamics acutely worsen after anesthesia induction prior to bypass in LVAD recipients, with worse hemodynamics among patients with preoperative temporary mechanical circulatory support. RV hemodynamics are not associated with anesthesia induction modality.

14 Abstract

Assessing the Course of Organ Dysfunction Using Joint Longitudinal and Time-to-Event Modeling in the Vasopressin and Septic Shock Trial

M Harhay[1], S Ratcliffe[2], J Russell[3]

1. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania
2. Division of Biostatistics, Department of Public Health Sciences, University of Virginia
3. Division of Critical Care Medicine, St. Paul's Hospital, University of British Columbia

Nonmortal septic shock outcomes, such as the Sequential Organ Failure Assessment [SOFA] score, are commonly analyzed clinical endpoints in pivotal sepsis trials. However, the assessment of longitudinal outcomes in critical care is confounded by high mortality rates and potentially differential effects of an intervention on mortality. Using data from the Vasopressin and Septic Shock Trial, a multicenter study of 779 septic shock patients from 27 ICUs in Canada, Australia, and United States, we compared the effects of vasopressin versus norepinephrine on the SOFA score over 28-days using a joint longitudinal and competing risks (for death and discharge) model. SOFA scores were calculated daily through discharge, death, or day 28. Arm was not directly associated with 28-day morality (35.0% vs. 39.3%; p=0.25). Changes in SOFA over time were associated with mortality (p<0.01), and SOFA trajectories were statistically different between the groups (Wald test χ^2= 42.0, p<0.01). Norepinephrine showed a more rapid decrease of SOFA compared to vasopressin over the first 4 days, with significant differences after 48 hours (SOFAVasopressin - SOFANorepinephrine = 1.17, 95% confidence interval: 0.46 to 1.87) and 72 hours (1.20, 0.42 to 1.98). The joint model is an accessible and statistically unbiased method to assess (it is simply a time-to-event and mixed-effects model that simultaneously inform each other) that could be used in future trials to assess the impact of interventions on the longitudinal progression of organ dysfunction, or other biomarkers, in patients with critical illness that are at risk of informative censoring bias due to mortality.

15 Abstract

Transfer Learning for Clustering Analysis from Single-Cell RNA-Seq Data

Jian Hu[1], Xiangjie Li[1], Gang Hu[1], Mingyao Li[1]

1. University of Pennsylvania

Recent development of single-cell RNA-seq technologies has led to enormous biological discoveries yet also introduced statistical and computational challenges. An important step in single-cell RNA-seq analysis is to cluster single cells into different cell types. Existing methods for cell type clustering suffer from low accuracy when the dataset has few cells or low sequencing depth. To overcome this limitation, we want to utilize knowledge from a relatively large training dataset to help cluster on a relatively small targeting dataset. As many single-cell studies are multi-year projects, it is appealing to transfer cell type knowledge learned from existing data to a new dataset generated in future years. Therefore, transfer learning suits perfectly for continuously generated data in these single-cell studies. We explored a transfer learning method to do cell clustering using single-cell RNA-seq data. In this method, one network developed for a task using training data is reused as the initial point for the clustering network on a second task using targeting data. To evaluate our method, we analyzed multiple human pancreas single-cell RNA-seq datasets. The clustering accuracy and efficiency were greatly improved if we utilize transferred information from training dataset compared to results without using the training data. We expect this approach will substantially improve the accuracy in cell type clustering, especially for small or low-quality datasets.

16 Abstract

Understanding Regression to the Mean Bias in the Context of Synthetic Controls

N. Illenberger[1], D. Small[1], P. Shaw[1]

1. University of Pennsylvania

To make informed policy recommendations from observational data, we must be able to discern true treatment effects from effects due to random noise and confounding. For longitudinal studies where data is split into pre- and post-treatment periods, techniques which match treated units to control units on pre-treatment outcomes, such as the synthetic control approach, have been presented as principled methods to account for confounding. However, we show how the use of synthetic controls or other matching procedures can introduce bias into estimates of the average treatment effect. In particular, we show how regression to the mean (RTM) bias can lead to substantially inflated type-I error rates and decreased power in typical policy evaluation settings. Through simulations, we compare the finite sample properties of common matching techniques and illustrate the need to account for RTM bias. Further, by conceptualizing observed treatment effects as the combination of true effects and effects due to RTM, we provide a novel correction for this bias which can eliminate bias and attain appropriate type-I error rates. As an illustration, we use our proposed correction to reanalyze data concerning the effects of California's Proposition 99, a large-scale tobacco control program, on statewide smoking rates.

17 Abstract

Reducing UV Exposure to Prevent Skin Cancer: Message Development and Testing

Amy Jordan[1], Amy Bleakley[1], DeAnn Lazovich[1], Andrew Strasser[1], Caroline La Rochelle[1], Karen Glanz[1]

1. University of Pennsylvania Prevention Research Center

Public Health Significance: Most skin cancers could be prevented if people adopted outdoor sun protection behaviors and avoided indoor tanning.

Purpose of Study: This study identified correlates of sun protection behaviors among adults aged 18 to 49 years and used this findings to develop and test messages and communication strategies for two categories of skin cancer risk behaviors: indoor tanning and outdoor sun exposure.

Methods: A mixed-methods strategy was applied to assess knowledge, attitudes and beliefs of adults aged 18 to 49 years in order to develop effective messages about skin cancer prevention. Specific target behaviors included avoidance of indoor tanning, use of sunscreen, protective clothing, and shade. The project blended field research in naturalistic settings, online panel theory-driven survey research, and cognitive, affective and biobehavioral testing of responses to health-promoting motivational and social marketing messages.

Results: Salient beliefs related to skin cancer prevention behaviors were identified for both outdoor sun exposure and indoor tanning. These findings were used to develop campaign messages that focused on skin cancer risk, appearance, and social norms in order to encourage sun protection behaviors in the target audience. Online experiments to test the message strategies suggest that for outdoor sun exposure, focusing on a single sun protection behavior at a time may be more effective than addressing multiple behaviors. Incorporating these findings into messages may increase the effectiveness of skin cancer prevention campaigns.

Conclusions: This study advances our understanding of strategies for skin cancer prevention campaigns.

18 Abstract

Airway Smooth Muscle-Specific Transcriptomic Signatures of Glucocorticoid Exposure

M Kan[1], C Koziol-White[2], M Shumyatcher[1], M Johnson[2], W Jester[2], RA Panettieri Jr[2], BE Himes[1]

1. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania

2. Rutgers Institute for Translational Medicine and Science, Rutgers University, State University of New Jersey, New Brunswick, NJ

Glucocorticoids, commonly used asthma controller medications, decrease symptoms in most patients, but some remain symptomatic despite high dose treatment. The physiological basis underlying glucocorticoid response, especially among asthma patients with severe, refractory disease is not fully understood. We sought to identify differences between fatal asthma and non-asthma donor-derived airway smooth muscle (ASM) cell transcriptomic response to glucocorticoid exposure, and to compare ASM-specific changes to those of other cell types. In cells derived from 9 fatal asthma and 8 non-asthma donors, RNA-Seq was used to measure ASM transcriptome changes after exposure to budesonide (100nM 24hr) or control vehicle (DMSO). Differential expression results were obtained for this dataset, as well as 13 publicly available glucocorticoid response transcriptomic datasets corresponding to 7 cell types. Specific genes were differentially expressed in response to glucocorticoid exposure: 7,835 and 6,957 in non-asthma and fatal asthma donor-derived ASM cells, respectively (adjusted p-value <0.05). Transcriptomic changes in response to glucocorticoid exposure were similar in fatal asthma and non-asthma donor-derived ASM, with enriched ontological pathways that included cytokine- and chemokine-related categories. Comparison of glucocorticoid-induced changes of the non-asthma ASM transcriptome to that of 6 other cell types showed that ASM has a distinct glucocorticoid response signature that is also present in fatal asthma donor-derived ASM cells.

19 Abstract*

Social Media Mining for Studying Patient-Reported Birth Defect Outcomes

A Klein[1], A Sarker[1], H Cai [1], D Weissenbacher[1], G Gonzalez-Hernandez[1]

1. Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania

Birth defects are the leading cause of infant mortality, but methods for studying their etiology (e.g., clinical trials, animal studies, pregnancy exposure registries) remain limited. To assess whether social media data could be used to observe pregnancies with patient-reported birth defect outcomes, we mined 432 million tweets posted by 112,647 users who have publicly announced their pregnancy on Twitter. To retrieve sparse tweets that mention birth defects, we developed a ruled-based, bootstrapping approach that relies on a lexicon, lexical variants, regular expressions, post-processing, and distributional properties. To identify a cohort for epidemiological analysis, inclusion criteria were tweets indicating that the user's child has a birth defect, and accessibility to the user's tweets during pregnancy. We manually annotated 16,822 retrieved tweets, with inter-annotator agreement of $\kappa = 0.79$ (Cohen's kappa). We analyzed 646 users who posted true positive tweets, and identified 195 of them who met the inclusion criteria. Congenital heart defects are the most common birth defects reported on Twitter, consistent with their relative prevalence in the general population. Based on an evaluation of 4,169 tweets retrieved using alternative text mining methods, the recall of the tweet-collection approach was 0.95. Our results suggest that social media mining can complement existing methods of birth defects research by providing (1) an opportunity to observe the periconceptional period and the early period of the first trimester, (2) a means of long-term follow-up after birth, (3) internal comparator groups, and (4) an opportunity to explore unknown risk factors.

20 Abstract

The Influence of Geographic Dispersion on Outcomes of Hospitalized Medicine Service Patients

Rachel Kohn[1-4], MO Harhay[3-5], B Bayes[2,4], H Song[3,6], SD Halpern[1-4], MP Kerlin[1-4], SR Greysen[1-4]

1. Department of Medicine
2. Center for Clinical Epidemiology and Biostatistics
3. Leonard Davis Institute of Health Economics
4. Palliative and Advanced Illness Research (PAIR) Center
5. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine at University of Pennsylvania
6. The Wharton School at the University of Pennsylvania

Background: Hospital wards place patients according to service (e.g., medicine) to improve efficiency. When hospitals reach high occupancy, this organization often breaks down and many patients become "geographically dispersed" to alternate wards.

Methods: Retrospective cohort study of medicine service patients in 3 University of Pennsylvania Health System hospitals, 2014-2015. The primary exposure was geographical dispersion. The primary outcome was length of stay (LOS). Secondary outcomes were discharge to skilled nursing facilities (SNFs) and in-hospital mortality. We excluded patients who switched between geographically dispersed and localized wards (1%). We performed quantile and logistic regression models and accounted for clustering by hospital. Covariates included age, gender, race, ethnicity, insurance type, body mass index, Elixhauser comorbidities, intensive care unit days, ward transfers, hospital utilization during the prior 12 months, medications, procedures, Centers for Medicare and Medicaid Services severity risk adjustment, admission source, average daily census for medicine services by hospital over a patient's hospitalization, and type of dispersed ward.

Results: The study population included 18,802 visits: median age 60 years, (IQR 45-74), 55% female, 61% black. Geographically dispersed patients (33% of total) had significantly longer LOS ($p<0.05$) at the median and above: 25th%=0.1 days, 95% CI 0.003-0.3, p=0.05; 50th%=0.2 days (0.05-0.4); 75th%=0.6 days, (0.4-0.9); 90th%=0.8 days (0.4-1.3). There were no differences in discharge destination (OR 1.0 [0.9-1.2]) or in-hospital mortality (OR 1.0 [0.5-2.0]).

Conclusions: As patients exceed median LOS, geographical dispersion is associated with increased LOS. Future studies are needed to confirm this finding and explore the underlying mechanisms of this association.

21 Abstract*

Learning Features from Longitudinal Data

William La Cava[1], Jason H. Moore[1]

1. Institute for Biomedical Informatics

Electronic health records (EHR) are a valuable resource that can be leveraged to improve patient care. Despite the potential for EHR data, current statistical and machine learning (ML) methods are limited in their capacity to learn from these data for a variety of reasons. A primary concern is that many commonly used methods do not natively handle mixed data types or longitudinal data collected at non-uniform intervals. To address this issue, we propose a tool that optimizes the data pre-processing pipeline by transforming longitudinal data into a tabular form for use with ML methods. The method uses an evolutionary computing technique to search for appropriate statistical transformations of the longitudinal predictors with the goal of maximizing classification accuracy and minimizing the complexity of the operators. As a result, a set of white-box, engineered features are produced. We first demonstrate the ability of this method to identify appropriate transformations on simulated data where the ground truth is known. We then apply the method to predict disease diagnoses in hospital patients using their clinical lab measure histories. The results show an improvement in predictive power for several target diseases in comparison to simple hand-engineered approaches.

22 Abstract

Scalable Automated Machine Learning and Applications on RNA-Seq Expression Data

T Le[1] , W Fu [1], J Moore [1]

1. Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania

TPOT is a Python Automated Machine Learning (AutoML) tool that uses genetic programming to help data scientists find the optimal machine learning pipelines for analyzing biomedical data. In the early implementations, like other AutoML tools, TPOT faces the challenges of long runtime, large computational expense as well complex pipeline with low interpretability, especially when analyzing big data which have recently become prevalent in many fields including biomedical research. We develop two novel features for TPOT, Template and Dataset Selector (DS), that help alleviate these issues. Specifically, Dataset Selector leverages domain knowledge while Template simplifies the pipeline structure to substantially reduce the computational expense. Together, these new features flexibly extend TPOT's application to biomedical big data analysis. We test the implementation of TPOT-DS integrated with Template in a computationally rigorous way. TPOT-DS significantly outperforms standard TPOT as well as a state-of-the-art tree-based machine learning method on realistic and challenging simulations that include complex network interaction effects. We also apply TPOT-DS to a real RNA-Seq study of major depressive disorder and identify data subsets of biological interest. Specifically, TPOT-DS repeatedly selects subset DGM-5 of which expression score has been previously shown to be associated with depression severity. With the new operators, our TPOT implementation is the first AutoML tool to offer the option of feature selection at the group level, making it computationally efficient and readily applicable to challenging datasets.

23 Abstract

(Abstract Withdrawn)

24 Abstract

Sparse Multiple Co-Inertia Analysis with Application to Integrative Analysis of Multi -Omics Data

Eun Jeong Min[1], Qi Long[1]

1. Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania

Multiple co-inertia analysis (mCIA) is a multivariate analysis method that can assess relationships and trends in multiple sets of data.

Recently it has been applied for integrative analysis of multiple high-dimensional -omics data.

However, the estimated loading vectors from the existing mCIA method are non-sparse, presenting challenges for interpreting analysis results.

We propose a novel sparse mCIA (smCIA) method that produces sparse estimates and a structured sparse mCIA (ssmCIA) method that further enables the incorporation of structural information among variables such as those from functional genomics.

Both the proposed methods achieve model estimation and variable selection simultaneously and enhanced feature selection performance and the sparse estimates that resolves interpretability problem.

Extensive simulation studies demonstrate the superior performance of the smCIA and ssmCIA methods compared to the existing mCIA.

We also apply our methods to the integrative analysis of transcriptomics data and proteomics data from a cancer study, the NCI-60 cancer cell lines data, to show the effectiveness of our proposed methods.

25 Abstract

Model Selection for Clinical Metabolomics: Comparing the Power of Different Optimization Approaches for Coronary Artery Disease Diagnosis Prediction

A Orlenko[1], D Kofink[2], FW Asselberg[2], J H Moore[1]

1. Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania
2. Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht University, the Netherlands

Model selection with AutoML tools is a field of a high interest in biomedical research as it provides an explanation of the complex behavior of the underlying processes in an agnostic manner. Our tool of choice – Tree-based pipeline optimizer (TPOT) employs evolutionary computations to select the most optimized machine learning (ML) model out of the pool of various possible combinations of data preprocessors and supervised ML algorithms. Here we use TPOT to predict angiographic diagnosis of coronary artery disease (CAD) in the Angiography and Genes Study. In addition, we provide a guideline for TPOT-based ML pipeline selection based on various clinical phenotypes and high-throughput metabolic profiles.

We performed a comparative analysis of TPOT-generated classification pipelines with selected ML classifiers, optimized with competitive grid search approach, to compare different CAD phenotypes: no CAD vs non-obstructive CAD and obstructive CAD (Setup 1); no CAD and non-obstructive CAD vs obstructive CAD (Setup 2).

TPOT produced classification ML pipelines that outperformed pipelines that were optimized with more traditional grid search approach across multiple performance metrics including balanced accuracy and precision-recall curve for both phenotypic setups (balanced accuracy of 0.79 for Setup 1 and balanced accuracy of 0.78 for Setup 2).

The power of agnostic model selection with AutoML tool TPOT has been demonstrated for CAD diagnosis prediction in clinical metabolomics study. TPOT generated pipelines that outperformed selected classifiers optimized with exhaustive grid search.

26 Abstract

Prevalence and Characterization of Yoga Mentions in the Electronic Health Record

N Penrod[1], S Lynch[1], S Thomas, J Moore[1]

1. University of Pennsylvania

There is a growing patient population using yoga as a therapeutic intervention, but little is known about how yoga actually interfaces with healthcare in a clinical setting. To characterize how yoga is documented at a large academic medical center and to identify clinician-derived and patient-derived use cases for yoga as therapy we designed a retrospective observational study in the electronic health record at Penn Medicine between November 2006-November 2016. Here we show that each year, yoga-containing notes were distributed among an increased number of patients, clinicians, and clinical service departments. During the study period, 30,976 unique patients, 2,398 unique clinicians, and 41unique clinical service departments were affiliated with yoga notes. A text-based classifier was built to separate the notes into three groups: clinician-documented yoga, clinician-recommended yoga, and other miscellaneous mentions of yoga. Statistical enrichment of primary diagnostic codes within these groups revealed 9 patient-derived therapeutic use cases of yoga and 11 clinician-derived therapeutic use cases of yoga. In total, we identified 14 unique medical conditions for which clinicians and/or patients view yoga as an effective treatment option. Our results contribute to a growing body of evidence about the role integrative medicine plays in the treatment and management of chronic diseases, and specifically points to an increasingly important role for yoga in healthcare.

27 Abstract

A Conceptual Model to Evaluate Disease-related Stigma and Access to Healthcare among Patients with Hepatitis C Virus Infection

M. Elle Saine[1,2], Julia E. Szymczak[1], Vincent Lo Re III[1,3]

1. Center for Clinical Epidemiology and Biostatistics, Center for Pharmacoepidemiology Research and Training, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania
2. Leonard Davis Institute of Health Economics, University of Pennsylvania
3. Department of Medicine, Perelman School of Medicine, University of Pennsylvania

Hepatitis C virus (HCV) infection has tripled since 2010, driven by increasing injection opioid use. HCV causes more deaths each year than any other infectious disease. Despite curative medications, most patients remain undiagnosed, not linked into care, and untreated. Disease-related stigma, a social process linking individual attributes to medical diagnoses, is associated with poor health outcomes; this may be intensified among patients who experience multiple forms of stigma (e.g. racism, substance use stigma). Several studies have called for conceptual models to assess the impact of stigma on access to healthcare. To address this need, we propose a conceptual model of the determinants and consequences of HCV-related stigma in access to healthcare.

We conducted semi-structured interviews among 20 patients with HCV infection who presented to care at five academic and community-based outpatient clinics across Philadelphia. We collected narrative data on patient experiences with HCV, social stigma, and healthcare.

We identified multiple intersecting stigmatized attributes, such as injection drug use, HIV coinfection, race, and poor HCV-related knowledge, which may amplify experiences of stigma among patients with HCV. As a result of HCV stigma, patients described social distancing by family and friends, often choose not to disclose their HCV status, and/or felt socially and emotionally isolated.

We developed a conceptual model linking HCV-stigma; stigmas of other attributes; and, the social barriers to HCV testing, care, and treatment adherence. This model can be used to guide research and interventions to improve patient access to care and health-related quality of life.

28 Abstract

Assessing qSOFA as a Predictor of In-ICU Infection and Non-Infection Outcomes in a Resource-Limited Setting in South Africa

SM Savarimuthu[1], RD Wise[2,3], C Cairns [2,3], NL Allorto[4], SD Halpern [5-8], GL Anesi[5-8]

1. Department of Medicine, Hospital of the University of Pennsylvania
2. Pietermaritzburg Department of Anaesthesia, Critical Care and Pain Management, Pietermaritzburg, South Africa.
3. Discipline of Anaesthesia and Critical Care, School of Clinical Medicine, University of KwaZulu-Natal, Durban, South Africa
4. Discipline of Surgery, School of Clinical Medicine, University of KwaZulu-Natal, Durban, South Africa.
5. Division of Pulmonary, Allergy, and Critical Care, Perelman School of Medicine, University of Pennsylvania
6. Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania,
7. Leonard Davis Institute of Health Economics, University of Pennsylvania,
8. Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania

The purpose of this study was to examine the accuracy of qSOFA in predicting ICU mortality in a low-resource setting. We performed a retrospective cohort study of patients admitted to Edendale Hospital ICU (South Africa) over 4 years, examining the association of qSOFA and SIRS criteria with in-ICU mortality using multivariable logistic regression models. Among 2,299 patients admitted to the ICU, observed in-ICU mortality was 15.3%. Among 1,272 (55.3%) patients with confirmed/suspected infection, 69.8% met qSOFA and 97.5% met SIRS definitions. In patients with infection, compared to a qSOFA score of 0/1, a qSOFA score of 2 was not associated with in-ICU mortality (OR = 1.29, 95% CI: 0.76-2.19, p = 0.34), but a qSOFA score of 3 was associated with an increased odds of in-ICU mortality (OR = 2.32, 95% CI 1.33-4.07, p = 0.003). Among patients without infection, qSOFA scores of 2 (OR = 4.22, 95% CI 1.81-9.82, p = 0.001) and 3 (OR = 10.40, 95% CI 4.15-26.08, p < 0.001) were associated with increased odds of in-ICU mortality; similar associations were observed in the trauma sub-population. SIRS was not statistically associated with in-ICU mortality for any subgroup. qSOFA discrimination for mortality was similar for patients with and without infection based on AUROC. In a resource-limited hospital, qSOFA was a strong predictor of in-ICU mortality for patients with and without infection and with trauma.

29 Abstract

Toward the Development of Dynamic Prediction Models for Clinical Applications

E Schnellinger[1], W Yang[1], M Harhay[1], S Kimmel[1]

1. University of Pennsylvania

Prediction models are used to make decisions in many areas of medicine. However, the current paradigm for developing these models is static: models are developed within a fixed derivation cohort, but applied to individuals outside this cohort (who may differ in clinical characteristics or disease risk). Furthermore, models are typically revised using a new derivation cohort, often yielding an entirely different model. Consequently, models are updated infrequently, and clinical decisions are made for years based on models with poor predictive accuracy. A new paradigm, dynamic prediction modeling (DPM), updates existing models by combining information used by the original model with new data collected from new patients. Yet in order to employ DPM in clinical practice, one must determine whether model updating provides meaningful improvements in risk prediction (acceptance threshold, α); how often to update the model (update interval, Δ); and how much importance to give to new versus old data (forgetting factor, ε). The goal of this ongoing project is to use statistical simulations to identify the appropriate thresholds for α, Δ, and ε when the true outcome rate or covariate coefficients change. Toward that end, we have simulated data informed by real-world post-transplant mortality among individuals who received lung transplantation in the United States during 2007-2015. Preliminary analyses suggest that comparing the parameters of the current prediction model to those from the true data generating model at baseline can aid in the identification of α. Future research will attempt to identify the appropriate thresholds for Δ and ε.

30 Abstract

Protective Equipment and Concussion in Women's and Men's Lacrosse: Findings from the Ivy League-Big Ten Epidemiology of Concussion Study

T Soya[1] , B D'Alonzo[1] , E Dorman[1] , K Arbogast[1] , D Wiebe[1]

1. University of Pennsylvania

Background: While both men's and women's lacrosse allow for stick checks, only male athletes are required to wear helmets. Women are required to wear only eye masks. We explore whether equipment differences by gender are accompanied with differences in mechanism of concussion.

Methods: We examine sport-related concussion in men's and women's lacrosse participating in the Ivy League-Big Ten Epidemiology of Concussion Study during 6 consecutive seasons (2013-2014 to 2018-2019), use of protective equipment, and mechanism of injury. Descriptive statistics calculated using STATA15.

Results: 180 concussions (n=72 women, n=108 men) occurred during the study period. Approximately one-third (n=63, 35.0%) were sustained in competition, and two-thirds (n=117, 65.0%) were sustained in practice. Reports from athletic trainers and narratives recorded in the database revealed that both men's and women's players wore required protective equipment over 93% of the time. The mechanism of concussion injury differed by sex: concussed women reported being hit by a stick (26.4%) or ball (37.5%) versus other mechanisms (36.1%) significantly ($p<0.001$) more frequently than did concussed men (stick=11.1% and ball=21.3% versus other=67.6%).

Conclusion: In this large sample of athletes, a significantly higher proportion of women's concussions were associated with hits to the head via ball or stick than men's concussions. These differences are likely primarily driven by the differential policy whereby men wear helmets per equipment requirement whereas women do not. Results demonstrate that concussions in women's lacrosse are associated with mechanisms that may be preventable by introducing helmets.

31 Abstract

Infertility and Mortality

NC Stentz[1], N Koelper[1], MD Sammel[1,2], KT Barnhart[1], S Senapati[1]

1. Reproductive Endocrinology & Infertility, University of Pennsylvania
2. Biostatistics, Epidemiology & Informatics, University of Pennsylvania

Importance: Infertility is a disease that affects 10% American reproductive-age women. The impact of infertility beyond the reproductive years is unknown.

Objective: To determine the association of infertility with all-cause and cause-specific mortality.

Methods: This secondary analysis of a multi-center RCT included women prospectively enrolled in the PLCO Cancer Screening Trial and followed 10-years for health-related outcomes and death. We examined the association of infertility history (inability to conceive for one year or greater) with all-cause and cause-specific mortality using disease risk score adjusted Cox-proportional hazard models.

Results: Infertile women are 12% more likely to die during the study period than the unexposed (AHR 1.12, 95%CI 1.05-1.19). The risk of death from diabetes is increased 71% in infertile women compared to the unexposed (AHR 1.71, 95%CI 1.15-2.56). The risk of death from cancer is increased 37% in infertile women at an otherwise low risk of cancer death compared to the unexposed (AHR 1.37, 95%CI 1.18-1.59). No differences are seen in the risk of death from endometrial/ovarian cancer. The risk of death from breast cancer is doubled in infertile women at an otherwise low risk of breast cancer death compared to the unexposed (AHR 2.36, 95%CI 1.62-3.43).

Conclusion: Infertility is a harbinger of morbidity and mortality. Infertile women are at an increased risk of all-cause, diabetes and cancer-related mortality. Consideration of infertility in health care maintenance presents an opportunity for screening and early intervention for long-term health outcomes.

32 Abstract*

Knowledge-Guided Bayesian Variable Selection in Non-Linear Support Vector Machines for Structured High-Dimensional Data

W Sun[1], C Chang[1], Q Long[1]

1. Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania

Support vector machine (SVM) is a popular classification method for the analysis of wide range of data including big data. Many linear SVM methods with feature selection have been developed through frequentist regularization, while the non-linear ones have a broader range of real applications. On the other hand, the importance of incorporating a priori known biological knowledge, such as gene pathway information which stems from the gene regulatory network, into the statistical analysis of genomic data has been recognized in recent years. In this article, we propose a new non-linear Bayesian SVM approach, based on the Gaussian process assumption, and guided by the knowledge on the graphical structure among predictors to perform feature selection. The proposed method uses a diagonal matrix with ones representing feature included and zeros representing feature excluded, and combines with the Ising prior that encourages group-wise selection of the predictors adjacent to each other on the known graph. MCMC sampling algorithm is used for Bayesian inference. The performance of our method is evaluated and compared with the standard kernel SVM methods in terms of prediction and feature selection in extensive simulation settings. In addition, our method is illustrated in the analysis of genomic data from a cancer study, demonstrating its advantage in generating biologically meaningful results and identifying potentially important features.

33 Abstract*

Genetic Analysis of Neuroblastoma in African-American Children

A Testori[1], Z Vaksman[1], S Diskin[1], J Maris[1], M Devoto[1]

1. Children's Hospital of Philadelphia

Neuroblastoma (NB), a pediatric cancer with a high degree of clinical heterogeneity, is rarer in African-American (AA) children compared to children of European descent. AA children with NB, however, more frequently develop the high-risk form of the disease and have associated lower overall survival. Using Genome Wide Association Studies (GWAS), we have identified several loci associated to NB in children of European descent, but only a few of them have been confirmed in AAs. Our expanded AA cohort of 674 cases and 3113 controls allowed us to thoroughly investigate the genetic susceptibility of NB in the AA population. Following high density genome-wide genotype imputation, we were able to confirm one major NB susceptibility gene (BARD1) in the AA population, which reached genome-wide significance in the subset of high-risk cases. Polygenic score analysis based on significance and estimates of SNP effect sizes from the European-American (EA) discovery GWAS, showed that the most significant score ($p = 2.3x10^{-14}$) included all SNPs with $p < 5.5x10^{-7}$ in the EA GWAS, and explained ~3% of NB risk variance in AAs. Other genetic analyses (including admixture mapping and haplotype association analysis) are in progress to test whether other NB susceptibility variants are located in regions of the genome that show different genetic ancestry in AA cases versus controls. These in particular may help explain susceptibility to developing the high-risk form of NB that disproportionately affects AA children with NB.

34 Abstract*

TAPAS: A Thresholding Approach for Probability Map Automatic Segmentation in Multiple Sclerosis

Alessandra M. Valcarcel[1], John Muschelli[2], Dzung L. Pham[3], Melissa Lynne Martin[1], Paul Yushkevich[4], Peter A. Calabresi[5], Rohit Bakshi[6,7], Russell T. Shinohara[1]

1. Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania
2. Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, 21287, United States
3. Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD, 20892, United States
4. Penn Image Computing and Science Laboratory (PICSL), Department of Radiology, University of Pennsylvania,
5. Department of Neurology, School of Medicine, Johns Hopkins University, Baltimore, MD, 21287, United States
6. Department of Neurology, Brigham Women's Hospital, Harvard Medical School, Boston, MA, 02115, United States
7. Department of Radiology, Brigham Women's Hospital, Harvard Medical School, Boston, MA, 02115, United States

White matter lesion volume is the most widely established magnetic resonance imaging (MRI) outcome measure in studies of multiple sclerosis (MS). Unfortunately, few approaches systematically determine the threshold employed to automatic segmentation probability maps; many methods use a manually selected threshold introducing human error and bias into the automated procedure. In this study, we propose and validate an automatic thresholding algorithm, Thresholding Approach for Probability Map Automatic Segmentation in Multiple Sclerosis (TAPAS), to obtain subject-specific threshold estimates for probability map automatic segmentation of T2 hyperintense white matter lesions. Using multimodal MRI, the proposed method applies an automatic segmentation algorithm to obtain probability maps. We obtain the true subject-specific threshold that maximizes Sørensen's-Dice coefficient (DSC) using a naive estimate of lesion volume generated from the probability maps. We model the subject-specific threshold on a naive estimate of volume using a general additive model. Using this model, we can predict a subject-specific threshold in data not used for training. We run a Monte Carlo-resampled split-sample cross-validation (100-fold) using two data sets.

By means of the proposed automated technique using Bland-Altman analysis, we found that volumetric bias associated with group-level thresholding is mitigated when applying TAPAS while increasing DSC.

The current study presents the first validated fully automated method for subject-specific threshold prediction.

35 Abstract

A Randomized Controlled Trial of Incentives Vs. Environmental Strategies for Weight Loss

K Volpp[1], P Shaw[1], K Hoffer[1], J Zhu[1], Q Huang[1], P Kwong[1], A Chung[1], R Choi[1], K Glanz[1]

1. University of Pennsylvania

Innovative approaches to weight loss and maintenance is an important public health priority. This study uses a web-based platform to enroll, communicate with, and track 344 participants from three diverse, obese employee populations. Participants were randomized to one of four arms to evaluate the comparative effectiveness of behavioral economic financial incentives and environmental strategies, separately and combined, in achieving initial weight loss and maintenance of weight loss over 24 months. Preliminary results show that at the 18-month primary endpoint, the incentive arm sustained weight loss relative to baseline. At 24 months, after 6 months without intervention, between-arm differences were not significant. There was an apparent weight loss relative to baseline maintained only in the combined arm.

The project objectives included:

- To assess the effectiveness of a daily lottery-based financial incentives, environmental strategies (ES), combined financial incentives and ES, relative to the control group, on cumulative weight loss over an 18-month period.
- To assess the cost effectiveness of the interventions relative to usual care

This poster provides an overview of the design and methods that were used in the study, which were presented at last year's DBEI Research day. Data collection was still ongoing at that time. This is a new poster updated to include preliminary results, conclusions, and policy implications.

36 Abstract

Social and Environmental Variables Obtained from Secondary Data Sources Explain Spatial Trends in Asthma Exacerbations Found in Electronic Health Record (EHR)-Derived Data

Sherrie Xie[1], Blanca E. Himes[1]

1. Department of Biostatistics and Epidemiology, University of Pennsylvania

Rationale: EHR-derived data is valuable to study phenotype and comorbidity relationships among real-life patients, but assessing the health impacts of social and environmental factors using EHR data alone is limited by the unavailability of relevant variables.

Methods: We obtained de-identified patient-level data for adult asthma encounters within the University of Pennsylvania Health System in 2014-2016. Residential geocodes were used to link clinical data to social and environmental data variables, including socioeconomic status (SES), crime, and traffic exposure. Cases were defined as asthma patients who had at least one asthma exacerbation, while controls had no exacerbations. Multivariable logistic regression models were used to identify factors associated with asthma exacerbations. Generalized additive models (GAMs) were used to perform spatial analysis and identify geographic regions associated with asthma exacerbations.

Results: Multivariable logistic regression on data for 3,661 patients (1,343 cases and 2,318 controls) found that, among the variables incorporated from external data sources, low neighborhood SES was significantly associated with asthma exacerbations. GAM analysis on data for 1,568 patients (570 cases and 998 controls) found a significant exacerbation hot spot ($p<0.01$) when the model included only EHR-derived covariates. However, GAM analysis found no hot spots when the model incorporated externally sourced variables.

Conclusions: Spatial trends in asthma exacerbation rates in Philadelphia determined from EHR data can be explained by differential patterns of neighborhood SES. Linking EHR-derived data with secondary sources of data on the social and physical environment provides a cost-effective means to understand health-related factors in real-life populations.

37 Abstract

Effective Strategies for Extending Relief-Based Feature Selection to Large Scale Feature Spaces

A Xu[1], S Bose[2], J Moore[1] R Urbanowicz[1]

1. University of Pennsylvania
2. Children's Hospital of Philadelphia

Feature selection is a ubiquitous element of data science. In the biomedical informatics domain, it is particularly important to develop feature selection methodologies that (1) scale to high dimensional feature spaces, (2) function even in the presence of limited sample sizes, (3) accommodate mixed variables types, (4) adapt to binary class, multiclass, and quantitative outcomes, (5) operate in the presence of missing data, (6) are sensitive to complex patterns of association (e.g. epistasis and genetic heterogeneity), (7) yield relative feature importance weights, and (8) are computationally efficient. Towards these ends we recently developed the ReBATE software package, implementing the novel MultiSURF algorithm and a family of existing Relief-based algorithms, demonstrating their success over a wide spectrum of simulated genetic association studies. Our recent work seeks to extend this package with a variety of existing 'wrapper' algorithms (i.e. TuRF, IterativeRelief, VLSRelief, and iVLSRelief) aimed in particular at scaling feature selection performance at detecting complex feature interactions to very large feature spaces. We also propose a novel hybrid rapper that combines the concept of iterative feature weight updates with a 'divide and conquer' approach championed by VLSRelief. We demonstrate the efficacy and compare the performance of these extensions within the ReBATE software over the course of a comprehensive simulation study that targets all of the aforementioned challenges.

38 Abstract*

Model-Based Phenotyping in Electronic Health Records with Data for Anchor-Positive Cases and Unlabeled Patients

Lingjiao Zhang[1], Xiruo Ding[2], Yanyuan Ma[3], Naveen Muthu[4], Imran Ajmal[5], Jason H. Moore[1], Daniel S. Herman[2], Jinbo Chen[1]

1. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania
2. Department of Pathology and Laboratory Medicine, University of Pennsylvania
3. Department of Statistics, Penn State University, Philadelphia, Pennsylvania, USA
4.Department of Biomedical and Health Informatics, University of Pennsylvania
5. Department of Pathology and Laboratory Medicine, University of Pennsylvania

Building phenotype models using electronic health record (EHR) data normally requires manually labeled cases and controls. Assigning labels is labor intensive and, for some phenotypes, identifying gold-standard controls is prohibitive. To facilitate comprehensive clinical decision support and research, we sought to develop an accurate EHR phenotyping approach that assesses its performance without a validation set.

Our framework relies on specifying an incomplete set of cases using an anchor variable that has perfect positive predictive value and sensitivity that is independent of predictors. We developed a novel maximum likelihood approach that efficiently utilizes data from anchor-positive and unlabeled patients to develop a logistic regression phenotyping model. Additionally, we describe novel statistical methods for estimating phenotyping prevalence and assessing model calibration and predictive performance measures. Theoretical and simulation studies indicate our method generates accurate predicted probabilities, leading to excellent discrimination and calibration, and consistent estimates of phenotype prevalence and anchor sensitivity. The method appears robust to minor lack of fit and the proposed calibration assessment can detect major lack of fit. We applied our method to EHR data to develop a preliminary model for identifying patients with diagnosed primary aldosteronism, which achieved an AUC of 0.99 and PPV of 0.8.

Our approach decreases labor-intensive manual phenotype labeling, which should enable broader model development and dissemination for EHR clinical decision support and research.

39 Abstract*

Identifying Signals of Potential Drug-Drug Interactions Involving Oral Anticoagulants

M Zhou[1], CE Leonard[1], CM Brensinger[1], WB Bilker[1], S Kimmel[1], T Hecht[1], S Hennessy[1].

1. Center for Pharmacoepidemiology Research and Training, Center for Clinical Epidemiology and Biostatistics, and Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine

Introduction: Drug drug interactions (DDIs) with oral anticoagulants are associated with increased risk of serious bleeding. Warfarin is susceptible to numerous DDIs while few studies have examined DDIs involving direct oral anticoagulants. We aim to identify medications that are most likely to increase the risk of serious bleeding when taken concomitantly with oral anticoagulants.

Methods: We conducted a high-throughput pharmacoepidemiologic screening study using OptumInsight Clinformatic Data Mart, 2000-2016. We performed self-controlled case series studies among adult oral anticoagulant users (warfarin, dabigatran, rivaroxaban, apixaban, and edoxaban) with at least one hospital presentation of serious bleeding. We identified all oral medications frequently co-prescribed with oral anticoagulants as potential interacting precipitants. Conditional Poisson regression was used to estimate rate ratios and 95% confidence intervals comparing precipitant exposed vs. unexposed time for each anticoagulant-precipitant pair. To control for within-person confounding by indication for the precipitant and to distinguish a DDI from a native effect of the precipitant on bleeding risk, we used pravastatin as a negative control object drug. Multiple estimation was adjusted using Semi-Bayes shrinkage.

Results: We screened 1,561 oral anticoagulant-precipitant drug pairs and identified 237 (15%) drug pairs associated with statistically significantly elevated risk of serious bleeding. Using pravastatin as the reference group, we identified 87 potential DDI signals for serious bleeding and 62(71%) of these were not documented in DDI knowledge databases Lexicomp and/or Micromedex.

Conclusion: We reproduced some previously documented DDIs, which demonstrated the validity of our approach. The newly identified potential DDI signals need be examined in future studies.

Notes

Perelman School of Medicine
University of Pennsylvania

DBEI & CCEB RESEARCH DAY

#2019ResearchDay

VISIT US at

www.dbei.med.upenn.edu
www.cceb.med.upenn.edu