First Annual Twin Cities ASA
FALL RESEARCH CONFERENCE
MAYO CLINIC, ROCHESTER, MN
Friday, October 30, 2015

SCIENTIFIC PROGRAM

ASA Traveling Short Course

Bayesian Methods and Computing for Evidence Synthesis and Network Meta-Analysis

Brad Carlin, University of Minnesota Division of Biostatistics

As the era of "big data" arrives in full force for health care and pharmaceutical development, researchers in these areas must turn to increasingly sophisticated statistical tools for their proper analysis. Bayesian statistical methods, while dating in principle to the publication of Bayes' Rule in 1763, have only recently begun to see widespread practical application due to advances in computation and software. This half-day short course reviews Bayesian methods, computing, and software, and goes on to elucidate their use in evidence synthesis and network meta-analysis (NMA). Broad application of these methods has been driven by an increased need for quantitative health technology assessment (HTA), especially comparative effectiveness research (CER). In particular, Bayesian methods facilitate borrowing of strength across treatments, trials, and outcomes (say, both safety and efficacy), as well as provide a natural framework for filling in missing data values that respect the underlying correlation structure in the data. We include descriptions and live demonstrations of how the methods can be implemented in BUGS, R, and versions of the BUGS package callable from within R.

Invited Speaker Session

Envelopes: Methods for Improving Efficiency in Multivariate Statistics

R. Dennis Cook, University of Minnesota School of Statistics

An envelope is a nascent construct for increasing efficiency in multivariate statistics without altering the traditional goals. Envelope estimation, which can produce efficiency gains equivalent to taking thousands of additional observations, is made possible by recognizing that the data may contain variation that is effectively immaterial to estimation or prediction. This informal notion leads to the central construct – an envelope – for enveloping the material information and thereby reducing estimative variation and improving inference.

Envelopes also link with some standard multivariate methodology. For instance, it was recently found that partial least squares regression depends fundamentally on an envelope at the population level, which opens the door to pursuing envelope estimators that can significantly improve upon partial least squares predictions.

We will begin with an intuitive introduction to response envelopes in the context of multivariate linear regression and then briefly describe some of their inner workings. This will be followed by a discussion of predictor envelopes and their connection to partial least squares. We will also describe how to extend the scope of envelope methods beyond linear models. The discussion will include several examples for illustration. Emphasis will be placed on concepts and their potential impact on data analysis.

Multivariate Analysis in Genetic Studies: Pleiotropy, Causal Modeling, and Disentangling Statistical Confusion

Daniel J. Schaid, Mayo Clinic Division of Biostatistics and Biomedical Informatics

Personalized medicine goes beyond simple disease status, capturing multiple traits that characterize a person's health status. The electronic medical record offers a rich source of laboratory measures, as well as health indicators. Yet, it is common to evaluate the association of a gene with each trait separately. For a long time, animal breeders have known what the human genetics community is beginning to discover: analyzing multiple correlated traits simultaneously can give greater power to detect underlying causal genes.

Yet, the statistical methods are somewhat disorganized, with "newly invented methods" unknowingly reinventions from existing methods. This talk will focus on recent developments of statistical models that are intended to simultaneously analyze how multiple traits are associated with a measured gene. Some methods will be familiar to biostatisticians, based on multivariate regression and canonical correlation, while other methods based on mixed models – the tools of animal breeders – will offer alternative insights of pleiotropy: multiple traits caused by a single gene. Furthermore, Bayesian multivariate causal models will be discussed, which attempt to partition direct and indirect effects of a gene on traits. By illustration of the some of the methods on existing data, strengths and limitations of the methods will be emphasized.

Design Considerations and Alternatives for Clinical Trials of Medical Devices

Ted Lystig, Medtronic

Graduate Student Posters

Hierarchical Bayesian Models for Understanding the Pharmacokinetics and Pharmacodynamics of Lorenzo's Oil

Cynthia Basu, Mariam Ahmed, James Cloyd, Richard Brundage, Reena Kartha, and Brad Carlin; University of Minnesota Division of Biostatistics

Lorenzo's Oil (LO) is a treatment available for X-linked adrenoleukodystrophy (X-ALD), a fatal neurodegenerative disease. Little has been done to establish its clinical efficacy. We analyze data on 116 male asymptomatic patients who were administered LO, including their fasting fatty acid profile (FAP). We adopt hierarchical Bayesian statistical approaches to understanding pharmacokinetics (PK) and pharmacodynamics (PD) of LO with respect to the FAP. We deal with the limitation of having one observation per cycle of the drug instead of multiple observation per cycle, as is usually used in PK-PD studies. Our action plan is to first link LO dose to the erucic acid concentration in the blood by PK modeling, and then link this concentration to a biomarker (C26, a very long chain fatty acid) by PD modeling. Next we design an adaptive Bayesian Phase IIa to estimate improvements in the biomarker from various LO doses accounting for possible toxicity and a Phase III study linking LO dose to actual improvements in health status.

A Temporal Analysis of Spectral Reflectance and Percentage Soil Carbon on Perennial Plants

Sabyasachi Bera, University of Minnesota School of Statistics

Perennial plants of various types were treated with various fungicides over the course of 6 years (2009-2014) on various plots and there spectral reflectance was measured in different times of the year. Also, amount of carbon in the soil was also measured after the growing season in 2014. Several parsimonious linear mixed models along with some useful graphical representation of the data were constructed to analyse the effect of different fungicides on plants, how they affect spectral reflectance and how soil carbon depends on spectral reflectance of the plants.

A Phenome-wide Scan to Detect Pleiotropic Effects of the Loss of Function R46L Variant in PCSK9

Brandon Coombes, University of Minnesota Division of Biostatistics

We performed a phenome-wide association study (PheWAS) of the missense mutation R46L (rs11591147) in PCSK9 with 1802 binary clinical phenotypes in 45,654 individuals (44.5% males) from the electronic MEdical Records and GEnomics (eMERGE) Network. A comprehensive range of phenotypes available in the electronic health record (EHR) was mapped using a hierarchical ICD-9 code-based PheWAS software (R statistical package PheWAS). A control set for each phenotype was constructed by selecting all patients that did not have the phenotype or closely related phenotypes. After standard quality control steps, R46L was imputed using IMPUTE2. Logistic regression with adjustment for age, sex, and race was used to test for associations between R46L and individual phenotypes. Bonferroni correction was applied to account for the testing of multiple phenotypes. We have also considered a gene-level PheWAS analysis of the PCSK9 gene by combining all of the available variant information with the aSPU test of Pan, 2014.

Enveloping the Aster Model

Daniel Eck, Charles Geyer, and Dennis Cook, University of Minnesota School of Statistics

Precise estimation of expected Darwinian fitness is a central component of life history analysis. Our methods provide precise estimation by incorporating general envelope model methodology into the aster modeling framework. The aster model serves as a defensible statistical model for distributions of Darwinian fitness. Envelope methodology reduces asymptotic variability by assuming a link between unknown parameters of interest and the asymptotic covariance matrices of their estimators. A novel envelope estimator is developed and used to obtain variance reduction. It is known both theoretically and in applications that incorporation of the preexisting general envelope model methodology and our novel envelope estimator reduces asymptotic variability. Our methods are tried on two simulated datasets. Variance reduction is obtained in the analyses.

A Bayesian Hierarchical Summary Receiver Operating Characteristic Model for Network Meta-analysis of Diagnostic Tests

Qinshu Lian and Haitao Chu, University of Minnesota Division of Biostatistics

In studies evaluating the accuracy of diagnostic tests, three designs are commonly used: (1) the crossover design; (2) the randomized design; and (3) the non-comparative design. Existing methods on meta-analysis of diagnostic tests mainly considered the simple cases when the reference test in all or none of the studies can be considered as a gold standard test, and when all studies use either a randomized or non-comparative design. Yet the proliferation of diagnostic instruments and diversity of study designs being used have boosted the demand to develop more general methods to combine studies with or without a gold standard test using different designs. In this paper, we extend the Bayesian hierarchical summary receiver operating characteristic model to network meta-analysis of diagnostic tests to simultaneously compare multiple tests under a missing data framework. It accounts for the potential correlations between multiple tests within a study and the heterogeneity across studies. In addition, it allows different studies to perform different subsets of diagnostic tests and provides flexibility on the choice of summary statistics. Our model is evaluated through simulations and illustrated using real data from deep vein thrombosis tests.

Alternative Measures of Between-Study Heterogeneity in Meta-Analysis: Reducing the Impact of Outlying Studies

Lifeng Lin, Haitao Chu, and James Hodges, University of Minnesota Division of Biostatistics

Meta-analysis has become a widely used tool to combine results from separate studies. The collected studies are homogeneous if they share a common underlying true effect size; otherwise, they are heterogeneous. A fixed-effects model is customarily used when the studies are homogeneous, while a random-effects model is used for heterogeneous studies. Assessing heterogeneity in meta-analysis is critical for model selection. Ideally, if heterogeneity is present, it should permeate the entire collection of studies, instead of being limited to a small number of outlying studies. Outliers can have great impact on the conventional measures of heterogeneity and the conclusions of a meta-analysis. However, no widely accepted guidelines exist for handling outliers. This article proposes several new heterogeneity measures. In the absence of outliers, the proposed measures are close to the conventional ones; in the presence of outliers, the proposed measures are less affected than the conventional ones. The performance of the proposed and conventional heterogeneity measures are compared theoretically, by studying their asymptotic properties, and empirically, using simulations and case studies.

Fast, Fully Bayesian Spatiotemporal Inference for fMRI Data

Donald Musgrove, University of Minnesota Division of Biostatistics

A Global Optimization Framework in the DC Setting

Daniel Prentice, Charles Geyer, and Xiaotong Shen, University of Minnesota School of Statistics

Difference of Convex (DC) functions represent a large and highly important class of functions in optimization and statistical model fitting. Previous global methods either have exponential time complexity or restrict themselves to a bounded parameter space, and faster local methods do not guarantee global optimality. We develop theorems that allow the globality of a proposed solution to be quickly verified even in an unbounded setting, and present a framework for fast global optimization in the DC space using existing local methods combined with our solution checking. Truncated Lasso Penalty (TLP) and other non-convex penalties have been suggested as superior alternatives to Lasso in applications like genome wide association studies, but that until now users of non-convex penalties could not be sure that they had global solutions. As an example we globally optimize a TLP regression problem.

A Bayesian Credible Subgroups Approach to Identifying Patient Subgroups with Positive Treatment Effects

Patrick Schnell, University of Minnesota Division of Biostatistics; Qi Tang, Walt Offen, Abbvie; Brad Carlin, UMN Biostatistics

Many new experimental treatments benefit only a subset of the population. Identifying the baseline covariate profiles of patients who benefit from such a treatment, rather than determining whether the treatment has a population-level effect, can substantially lessen the risk in undertaking a clinical trial and expose fewer patients to treatments that do not benefit them. The standard analyses for identifying patient subgroups that benefit from an experimental treatment either make separate marginal inferences on each individual, which raises multiplicity issues, or focus inappropriately on the presence or absence of treatment-covariate interactions. We propose a Bayesian credible subgroups method to identify two bounding subgroups for the benefiting subgroup: one for which it is likely that all members simultaneously have a treatment effect exceeding a specified threshold and another for which it is likely that no members do. We illustrate the approach using data from an Alzheimer's disease treatment trial and conclude with a discussion of the advantages and limitations of this approach to identifying patients for whom the treatment is beneficial.

Output Analysis for High-Dimensional Markov Chain Monte Carlo

Dootika Vats, University of Minnesota School of Statistics; James Flegal, UC–Riverside; Galin Jones, UMN Statistics

Markov chain Monte Carlo (MCMC) methods produce a correlated sample in order to estimate several (many) unknown expectations of a target distribution with a vector of sample means. Ensuring that this procedure is reliable requires assessment of the Monte Carlo error. However, the multivariate nature of the estimation process has been ignored in the MCMC literature. We present multivariate estimators of the Monte Carlo error and its implementation in an R package. These estimators allow for joint confidence regions for parameters and for the application of sequential stopping rules to determine termination for the Markov chain. We implement our methods on a Bayesian probit regression model.

WITHDRAWN: ~~*An Adaptive Association Test for Microbiome Data*~~

Chong Wu, University of Minnesota Division of Biostatistics

~~There is an increasing interest in investigating how the compositions of microbial communities are associated with various risk factors like environmental exposures. Distance-based analysis and microbiome regression based kernel association test (MiRKAT) are two popular methods for such analyses. A proper choice of a phylogenetic distance is critical for the power of these methods. However, existing phylogenetic distance metrics are designed without accounting for differential information contents with various microbial lineages: given that not all microbial lineages are expected to be associated with the risk factor of interest, using all the lineages in distance calculations introduces noises, leading to power loss in the subsequent association testing. We propose a class of microbiome based sum of powered score (MiSPU) tests based on a newly defined generalized taxon proportion that combines observed microbial composition information with phylogenetic tree information. Different from the existing methods, a MiSPU test is based on a weighted score of the generalized taxon proportion in a general framework of regression, upweighting more likely to be associated microbial lineages. Our simulations demonstrated that one or more MiSPU tests were more powerful than MiRKAT while correctly controlling type I error rates. An adaptive MiSPU (aMiSPU) test is proposed to combine multiple MiSPU tests with various weights, approximating the most powerful MiSPU for a given scenario, consequently being highly adaptive and high powered across various scenarios. We applied MiSPU and aMiSPU to a throat microbiome dataset, showing that microbial communities were associated with the smoking status while adjusting for potential confounders. The proposed methods are available in the R package~~ `MiSPU`.

Powerful Association Testing via Accounting for Genetic Heterogeneity

Zhiyuan Xu and Wei Pan, University of Minnesota Division of Biostatistics

Genome-wide association studies (GWASs) have confirmed the ubiquitous existence of genetic heterogeneity for common disease: multiple common genetic variants have been identified to be associated, while many more are yet expected to be uncovered. On the other hand, the single SNP-based trend test (or its variants) that has been dominantly used in GWASs is based on contrasting the allele frequency difference between the case and control groups, completely ignoring possible genetic heterogeneity. In spite of the widely accepted notion of genetic heterogeneity, we are not aware of any previous attempt to apply genetic heterogeneity-motivated methods in GWAS. Here, to explicitly account for unknown genetic heterogeneity, we applied a mixture model-based single SNP test to the WTCCC GWAS data with traits Crohn's disease, bipolar disease, coronary artery disease and type 2 diabetes, identifying much larger

numbers of significant SNPs and risk loci for each trait than those of the popular trend test, demonstrating potential power gain of the mixture model-based test.

Clustering for Personalized Prediction

Fan Yang and Xiaotong Shen, University of Minnesota School of Statistics

We build a model to allow personalized prediction for different individuals on a large amount of items based on both user features and item features, as in a recommender system. User and item "preferences" are clustered through supervised learning by modeling the observed response with a gaussian distributed regression model. Besides mean parameters, correlation structure of the response variable is also modeled. Fusion type penalties are applied to identify similar users and items. Simulation results show our model performs better than the popular matrix decomposition methods.

Combining Non-randomized and Randomized Data in Clinical Trials Using Commensurate Priors

Hong Zhao and Brad Carlin, University of Minnesota Division of Biostatistics

Historically, only randomized clinical trials (RCTs) have been recognized as the gold standard for evaluating the efficacy or safety of a therapeutic intervention. Although RCTs have reliable internal validity, they often are restricted to well-defined groups. By contrast, observational studies (OSs) may have better generalizability, due to a broader subject pool. However, OSs may suffer from selection bias without proper adjustment for potential confounders. Therefore, combining RCTs and OSs in research synthesis is often criticized due to the limitations of OSs. Recent research suggests that systematic reviews of treatment effects should not be restricted to specific study types in all cases. In this work, we develop hierarchical Bayesian approaches that combine data from all sources simultaneously while explicitly acknowledging the differences in designs. Specifically, we are proposing a novel two-step approach to first match non-randomized data using propensity score method, and then use commensurate priors to integrate information from the matched non-randomized studies with data from RCTs to an extent that depends on estimates of the commensurability of the data sources. We apply the proposed framework to a recent HIV clinical trial, and investigate the operating characteristics of our methods via simulation. Our findings elucidate the extent to which well-designed non-randomized studies can complement RCTs for improved clinical decision making.

Causal vaccine effects on binary post-infection outcomes: a Bayesian approach vs. the maximum likelihood method

Jingcheng Zhou, Haitao Chu, University of Minnesota Division of Biostatistics; Michael G. Hudgens, UNC–Chapel Hill; M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center

To estimate causal effects of vaccine on post-infection outcomes, Hudgens and Halloran (2006) defined a post-infection causal vaccine efficacy estimand VE_I based on the principal stratification framework. They also derived closed forms for the maximum likelihood estimators of the causal estimand under some assumptions. Extending their research, we propose a Bayesian approach to estimating the causal vaccine effects on binary post-infection outcomes. The identifiability of the causal vaccine effect VE_I is discussed under different assumptions on selection bias. The performance of the proposed Bayesian method is compared with the maximum likelihood method through simulation studies and two case studies – a clinical trial of a rotavirus vaccine candidate and a field study of pertussis vaccination. For both case studies, the Bayesian approach provided similar inference as the frequentist analysis. However, simulation studies with small sample sizes suggest that the Bayesian approach provides smaller bias and shorter confidence interval length.

Undergraduate Student Posters

Integration of Path Analysis to Bayesian Networks for Modeling Latent Variables

Charles Cain, University of Minnesota–Morris

Causal Inference has been a large area of study in Biostatistics and Epidemiology to help explain what specifically causes certain diseases and conditions. Bayesian networks use conditional probabilities in order to make these causal inferences. The simplest types of Bayesian Networks are networks made up of observable discrete or discretized continuous variables. However, many authors have shown advantages of continuous variables in Bayesian Networks over their discretization. Also, the use of latent variables or hidden variables have been used as a factor in explaining variables inside the Bayesian Network, discrete or continuous. This research looks at latent variables as a weighted sum of observed variables much like in Path Analysis or Structural Equations. We then use these modeled latent variables as continuous variables in a Bayesian Network. As an example we will look at a Bayesian Network of the causation of Diabetes

using data from NHANES and modeling a latent variable, "Lack of Physical Activity", as a weighted sum of variables in the data.

Statistical Analysis of K–12 Assessment Data from the Northfield Public School District

Doug Carmody, Miranda Tilton, and **Gwen Vargas**, St. Olaf College

MHEDS R-Project: Course Development, Spatio-Temporal Data Visualization and Phylogenetic Classification

Akanksha Dua, Cory Stern, Kwangu Yeo, Yuqi Ren, Macalaster College

As Embedded Data Scientists, our goals included implementing R statistical software into a course at Macalester College, and also provide assistance in various research projects using techniques in R. We analyzed topographical and evolutionary data and also made statistical concepts accessible to students by creating video podcasts and interactive learning platforms in R.

Modeling Latent Variables with Directional Dependence

Humza Haider, University of Minnesota–Morris

Being able to understand and model multivariate dependence based on direction is a challenge that many theoretical and applied researchers face. Directional dependence can help fields such as finance, biostatistics, economics, and bioinformatics. Latent variables are also present among many fields. Latent variables often appear when some unmeasurable variable needs to be used as a predictor. Latent variables are present in many studies and so their application to research is obvious. Being able to understand how a latent variable is formed can further increase the ability to analyze the latent variable in respect to the study. This paper uses the example of a physical inactivity as a latent variable in trying to model diabetes in a logistic regression. We will apply two different methods of directional dependence on latent variables, one which maximizes the correlation between all the variables and another which maximizes the correlation between a dependent variable and a weighted sum of the latent variable variables. These models will then be compared to an undirected latent variable approach and a model without a latent variable. The data is taken from the NHANES (National Health and Nutrition Examination Survey) data set.

Hospital Coordination and Medical Mistakes

Nick Nooney, St. Olaf College

A Nonparametric Look at Self-Esteem Development

Mark Ruprecht, University of Minnesota

Comparison of Non-Parametric and Parametric Methods for Testing Changes in Likert-type Scale Assessments

Michael Shyne, Metropolitan State University

A test/control design study on self-care instruction methods was conducted with students in a graduate level nursing theory class. The outcomes were measured via self-care assessments consisting of Likert-type scale questions completed near the beginning of the term and again near the end. The question of interest was whether the test group assessment score difference ("post" minus "pre") increased more than the control group assessment score difference for seven score categories.

Changes from "pre" assessments to "post" assessments were compared for each category score and by group using both non-parametric and parametric methods. Test group scores increases were greater for every category than control group score increases, though the difference was significant for only nutrition and spiritual growth (median increase difference (test – control) = 0.167; Wilcoxon signed rank-test p=0.035 and 0.222; p=0.031 for nutrition and spiritual growth, respectively). Although, the difference was not a continuous, normally distributed variable, parametric methods produced similar results (p=0.035 and 0.027, for nutrition and spiritual growth, respectively).

Ambiguity in Congressional Campaign Communication: Assessing Unclear Political Statements Using Supervised Learning Methods

Jack Werner, St. Olaf College

Meta-analysis on Second Language Acquisition

Jiahao Zhang, St. Olaf College

This study reviews and analyzes the present state of research on the function of the elicited imitation task (EIT) in measuring adult second language acquisition. The effect size of meta-analysis shows a statistically significant correlation between EIT and other proficiency tests. However, the 46 studies included contained a variety of designs and methodologies, showing lack of agreement about which test design is ideal. Due to the variation in what researchers chose for comparison with EIT, six subgroups were formed. Finally, based on observation of the current research, recommendations are made for future research.

Other Posters

Adaptive Gene- and Pathway-Trait Association Testing with GWAS Summary Statistics

Il-Youp Kwak, University of Minnesota Division of Biostatistics

In spite of the tremendous success of genome-wide association studies (GWASs), due to small effect sizes of the majority of causal SNPs for complex and common disease, large sample sizes are always needed. As alternatives, gene- and pathway-based analyses have been proposed. A statistical challenge is that, due to unknown true association patterns, there is no uniformly most powerful test.

Pan et al. (2014) proposed a data-adaptive (aSPU) approach based on estimating and selecting the most powerful test among a class of sum of powered score (SPU) tests, which cover several popular tests as special cases. Furthermore, Pan et al, (2015) extended the methodology to pathway analysis (aSPUpath). In particular, two parameters are introduced such that the test is adaptive at both the SNP and gene levels.

However, the two adaptive tests for gene- and pathway-trait associations are only applicable to the case with individual-level genotype and phenotype data. It is often difficult to obtain access to individual-level data. We propose extending the two adaptive tests to the case with only summary statistics for individual SNPs, demonstrating their applications to a meta-analyzed GWAS dataset.

The methods are available in R package, `aSPU`.