CONTROL ID: 3276323

TITLE: A latent class based imputation under Bayesian quantile regression framework for longitudinal medication usage data with missing values

ABSTRACT BODY:
Abstract Body: Evaluating the association between diseases and the longitudinal pattern of pharmacological therapy has become increasingly important. However, in many longitudinal studies, self-reported medication usage data collected at patients' follow up visits could be missing for various reasons. These pieces of missing or inaccurate/untenable information complicate determining the trajectory of medication use and its complete effects for patients. Although longitudinal models can deal with specific types of missing data, inappropriate handling of this issue can lead to a biased estimation of regression parameters especially when missing data mechanisms are complex and depend upon multiple sources of variation. We propose a latent class based multiple imputation approach using a Bayesian quantile regression that incorporates cluster of unobserved heterogeneity for medication usage data with missing values. Findings from our simulation study indicate that the proposed method performs better than traditional imputation methods under certain scenarios of data distribution. We also demonstrate applications of the proposed method to real data obtained from the longitudinal cohort study that assesses a trajectory of medication usage and its association with disease progression, while self-reported medication usage data are incomplete during follow-up in the cohort.

AUTHORS/INSTITUTIONS: M. Lee, U of Texas McGovern Medical School, Houston, Texas, UNITED STATES
Abstract Body: Covariate-adjusted response-adaptive (CARA) designs use the available responses to skew the treatment allocation towards the best treatment found during a clinical trial, for a patient's covariate profile. Extensive research on CARA designs with survival responses assumed patients failing for a given cause, where failures for other causes are considered to be censored. In medical research, survival responses are often classified as failure from multiple causes. For example, in oncology trials monitoring deep molecular response, deaths often occur due to QT prolongation. In such cases, a competing event precludes the occurrence of the main event. Ignoring such information during the design stage may severely bias the results for treatment comparisons.

To make CARA designs more applicable in clinical trials with multiple causes of failure, they are developed by avoiding any distributional assumptions on the survival responses, but only assuming the proportionality of the subdistribution hazards between the two treatment arms. Proportionality of the cumulative incidence functions has also been considered to have a direct interpretation about survival probabilities for the main failure type. The proposed designs are based on biased coin procedures, with a bias towards the better treatment arm. These are the doubly-adaptive biased coin design (DBCD) and the efficient randomised adaptive design (ERADE). They achieve this ethical objective of a trial without compromising on the statistical power for treatment comparison. The treatment allocation proportions for these designs converge to the expected target values, which are functions of the sequentially estimated Fine and Gray regression coefficients. The asymptotic properties of these designs are extensively studied. Unlike the DBCD which is a continuous allocation rule, the asymptotic property of the discrete ERADE is established by introducing a stopping time of a martingale process.

Simulation results show that the ERADE is preferable to the DBCD when the main aim is to have an ethical design with minimum variance and maximum power of the Wald test for treatment comparison. The proposed methods are suitable alternatives to balanced designs in terms of power. They treat more patients in the trial with the best treatment, making them more ethical than the balanced designs. An existing clinical trial is re-designed using these methods.

Authors/Institutions: A. Mukherjee, Statistical Sciences, PPD International, Bangalore, Karnataka, INDIA
Abstract Body: The delicate balance of the microbiome is implicated in our health and is shaped by external factors. Therefore, understanding the mediating role of the microbiome in linking external factors and our health conditions is crucial to translate the microbiome research into therapeutic and preventative applications. We introduce a sparse compositional mediation model for binary outcomes under potential outcomes framework to estimate and test the causal mediation effect utilizing the compositional algebra defined in the simplex space and a linear zero-sum constraint on regression parameters.

Authors/Institutions: M.B. Sohn, Biostatistics and Computational Biology, University of Rochester, Rochester, New York, UNITED STATES|J. Lu, H. Li, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES|
Non-normality is a common phenomenon in data from agricultural and biological research, especially in molecular data (for example; -omics, RNAseq, flow cytometric data, etc.). For over half a century, the leading paradigm called for using analysis of variance (ANOVA) after applying a data transformation. The introduction of generalized linear mixed models (GLMM) provides a new way of analyzing nonnormal data. Selecting an apt link function in GLMM can be quite influential, however, and is as critical as selecting an appropriate transformation for ANOVA. In this paper, we assess the performance of different parametric link families available in literature. Then, we propose a new estimation method for selecting an appropriate link function with a suitable variance function in a quasi-likelihood framework. We apply these methods to a proteomics data set, showing that GLMMs provide a very flexible framework for analyzing these kinds of data.

AUTHORS/INSTITUTIONS: W. Malik, H. Piepho, Biostatistics Unit, University of Hohenheim, Stuttgart, Stuttgart, GERMANY
Nonlinear Mixed Effects Modeling of Height Growth from Early Childhood to Adult Age in Longitudinal Data

Settings: Model Comparison and Prediction

Abstract Body: Modeling human physical growth plays a vital role to examine and define growth trajectories related to health and wellbeing. The primary objective of this study is to model height growth from early childhood to adult age using nonlinear mixed models. Data used in this study belong to Young Lives Ethiopia, older cohort study where 889 children followed from age 90 to 240 months. Nonlinear mixed-effects models in structural and nonstructural frameworks have been applied to height measurements observed over time. Height growth velocity was the peak in females than males. The rate of change in the childhood period, during puberty period and time to experience puberty, had a negative correlation with adult height. Children in different regions had different Adult height, a rate of change before and during puberty, initial height and time at peak velocity. Age at peak height and take-off velocity in males was 24 months late compared to that of females. SITAR and Preece-Baines 1 (PB1) model best fitted and captured within and between the individual growth of females and males, respectively. Generally, the SITAR and the PB1 model had best-fitted height growth of children from early childhood to adult age and the PB1 model better-predicted height growth beyond data points.

Authors/Institutions: D.D. Debeke, Statistics, Hawassa University, Hawassa, SIDAMA, ETHIOPIA, ETHIOPIA
Abstract Body: Large amounts of longitudinal health records are now available for dynamic monitoring of the underlying processes governing the observations. However, the health status progression across time is not typically observed directly: records are observed only when a subject interacts with the system, yielding irregular and often sparse observations. This suggests that the observed trajectories should be modeled via a latent continuous-time process potentially as a function of time-varying covariates. We develop a continuous-time hidden Markov model to analyze the longitudinal data accounting for irregular visits and different types of observations. By employing a specific missing data likelihood formulation we can construct an efficient computational algorithm. We focus on Bayesian inference for the model: this is facilitated by an expectation-maximization algorithm and Markov chain Monte Carlo. Simulation studies demonstrate that these approaches can be implemented efficiently for large data sets in the fully Bayesian setting. We apply this model to a real cohort where patients suffer from chronic obstructive pulmonary disease with the outcome being the number of drugs taken, using healthcare utilization indicators and patient characteristics as covariates.

AUTHORS/INSTITUTIONS: Y. Luo, Mathematics and Statistics, McGill University, Montreal, Quebec, CANADA
ABSTRACT BODY:

Abstract Body: BACKGROUND

Kenya, just like other soil transmitted helminthes (STH) endemic countries, has been conducting regular treatment program for the last five years among school aged children as a way to reduce STH infections burden in the country. However, the point of interruption of transmission of these infections still remains unclear. In the current study, we analyzed an age structured mathematical model to predict the point of interruption of these infections in Kenya. The main objective was to develop and analyze an age structured model of the STH population dynamics under a regular STH treatment program to determine infection transmission rate, the point of infection interruption, and the optimal interpulse treatment interval sufficient to achieve STH infections elimination in Kenya.

METHODS

The study utilized age structured model of the STH population dynamics under a regular treatment program. The model was applied to two main age groups: school age children (5-14 years) and adult populations (>15 years) and investigated the potential for STH elimination with finite rounds of treatment while allowing the STH distribution to change dynamically as a function of treatment frequency and treatment coverage. The model was verified using a five year field data from the National School Based Deworming Program (NSBDP) for all the three main STHs; Ascaris lumbricoides, Trichuris trichiura and hookworms.

RESULTS

The model behaviour demonstrated convincingly an accurate predictions of prevalence and mean intensities of infections during and after treatment rounds in each of the age groups. The model indicated that the benefit derived from the regular treatment increases non-linearly with the treatment rounds and coverage. Additionally, it depicted that for elimination to be achieved within a shorter time period in the general population and within each age group, higher treatment coverage and biannual treatment rounds are more effective.

CONCLUSION

The model captured the dynamics of the STH burdens in vulnerable populations under regular treatment program as elimination is approached. It aided in examining the role of age structure to the persistent STH infections in Kenya. As a result of these findings, we aim to advise the STH control programs on the right mix of strategies needed to achieve faster elimination of the STH infections in Kenya.

AUTHORS/INSTITUTIONS: C. Okoyo, N. Owuor, School of Mathematics, University of Nairobi, Nairobi, KENYA|C. Okoyo, C. Mwandawiro, Eastern and Southern Africa Centre of International Parasite Control, Kenya Medical Research Institute, Nairobi, KENYA|G. Medley, Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, UNITED KINGDOM|
Considering the problem of identifying subgroup in a randomized clinical trial with respect to survival time, we present an analysis strategy to find subgroup of enhanced treatment effect. We fit univariate accelerated failure time (AFT) models with covariate-treatment interactions to identify predictive covariates. The false discovery rate is controlled by Benjamini-Hochberg procedure. Then a composite score conversion is employed to transform the set of identified covariates for each patient into a univariate score. To classify patient subgroups, a change-point algorithm is applied to searching for the threshold cutoff instead of using the median. Moreover, we adopted a biomarker adaptive design to check whether the treatment effect exists within certain subgroup. The simulation results show that the change-point method is remarkably superior to the median cutoff particularly when the subgroup sizes vary considerably. Furthermore, the 2-stage adaptive design has good power properties in detecting treatment effect while the type I error is generally controlled. As an illustration, we apply the proposed methods to an AIDS study. In conclusion, when the sample size is sufficient and the censoring rate is mild, the AFT model combined with change-point algorithm performs well in identifying subgroup.
ABSTRACT BODY: Neyman-Pearson’s theory of testing statistical hypotheses is mathematically beautiful and exclusively illustrated in most statistical text books. But its misuse in practice is not negligible. In 2016, the American Statistical Association released a statement against the misuse of statistical significance and P-values, and in March, 2019, Amrhein, V., Greenland, S., McShane, S., and more than 800 signatories called for the entire concept of statistical significance to be abandoned. We appreciate it, in particular, in Biostatistics, the calling for giving up of the teaching of Neyman-Pearson type testing statistical hypotheses; but coming back to Fisher’s statistical inference that instructs to consider whether the observed effect is practically meaningful if P-value $\leq 0.05$ and if meaningful, then decide it significance; and not to consider the data anymore if P-value $> 0.5$. If this is the case, the two important topics in statistics, one is the sample size determination in designing studies and the other is the optimality of test statistics used to compute P-values could drift on the air, since those topics were developed in the framework of Neyman-Pearson type tests of testing statistical hypothesis.

We show in this paper that the two topics can be settled without using the theory of Neyman-Pearson type test, but by introducing the concept of reproducibility of statistical test results based on P-value. Here, the reproducibility is defined as follows; suppose that the same test is repeated again as the first test, independently to the first test, if and only if the first test results significance. If the same result is obtained as the first test, we call the statistical test result has reproducibility. The reproducibility is natural and directly appeals to scientists whose biggest concern is the reproducibility of their findings.

AUTHORS/INSTITUTIONS: T. Yanagawa, Biostatistics Center, Kurume University, Kurume, JAPAN|K. Omae, Department of Clinical Biostatistics, Kyoto University, Kyoto, JAPAN
Background: In order to account for correlated count data with excess zeros, we use variational approximation multivariate latent generalized linear model. We performed two different simulation based on level species and genus with Poisson and negative binomial to subject-specific interpretations. Methods: In this work, we use variational approximation to estimate parameters in multivariate latent generalized linear model. Otherwise, overdispersed a count outcome exhibiting many zeros, in excess of the amount expected under sampling from a Poisson distribution. Results: Through simulation studies, species counts follow negative binomial, and genus counts follow Poisson distribution and the performance of this method evaluated by AIC, AICc, and BIC. Conclusion: While these two sets of latent class parameters might be meaningful in certain species counts and genus counts.

Abstract Body: Over the last two decades, advanced methods have been developed to model the spatio-temporal evolution of real-world epidemics. Tracking and forecasting the full spatio-temporal evolution of an epidemic can help public health officials to plan their emergency response and health care.

We present advanced methods of spatial data assimilation to epidemiology, in this case to the ebb and flow of Ebola across the landscape of northeastern provinces in the Democratic Republic of Congo (DRC).

Data assimilation is a general Bayesian technique for repeatedly and optimally updating an estimate of the current state of a dynamic model. We present a stochastic spatial Susceptible-Vaccinated-Exposed-Infectious-Recovered-Dead (S-V-E-I-R-D) compartmental model to capture the transmission dynamics and the spatial spread of the ongoing Ebola outbreak in the DRC.

In this application the machinery of data assimilation acts to integrate incoming weekly incidence data into a fully spatial population model of Ebola transmission, within a Bayesian framework for the tracking process. For the current outbreak in DRC we use registered data (province-wide weekly counts of total Ebola cases and confirmed dead) from the World Health Organization (WHO) situation reports.

Our simulations show good correspondences between the stochastic model and the available sparse empirical data. A comparison between weekly incidence data set and our SVEIRD model coupled with Bayesian data assimilation highlights the role of a realization conditioned on all prior data and newly arrived data. In general, the SVEIRD model with data assimilation gives a better fit than the model without data assimilation for the same time period.

Our analyses may shed light more broadly on how the disease spreads in a large geographical area with places where no empirical data is recorded or observed.

The analysis presented herein can be applied to a large class of compartmental epidemic models. It is important to remember that the model type is not particularly crucial for data assimilation, the Bayesian framework is the key. Data assimilation neither requires nor presupposes that the model of the infectious disease be in the family of S-I-R compartmental models. The projected number of newly infected and death cases up to March 1, 2020 are estimated and presented. We provide a discussion and interpretation of our results.

Authors/Institutions: A. Krishnamurthy, N. Hijazi, G. Dhaliwal, Mathematics and Computing, Mount Royal University, Calgary, Alberta, CANADA|L. Cobb, Mathematical and Statistical Sciences, University of Colorado Denver, Denver, Colorado, UNITED STATES|
Abstract Body: Medical research now tends to a more personalized approaches including diagnosis and treatment. This shift of perspective is often not reflected in the data analysis step. We here present an approach based on a mechanistic mathematical model of acute myeloid leukemia (AML). Our primary aim is to develop and validate personalized models to improve therapy decision making and prediction of treatment success. The basic ODE model was published by Stiehl et al. (2014) and a further analysis with respect to combination chemotherapy treatment was published by Banck and Görlich (2019), recently. For our analysis we use individual patient data from the AMLCG2008 trial (NCT01382147, Braess et al. 2018). Patients were treated with a combination chemotherapy, i.e. the S-HAM protocol. The ODE model was adapted to implement the S-HAM treatment course. Personalization is introduced into the model by fitting a subset of model parameters on individual patient data. The main outcome is the leukemic burden i.e. percentage of blast cells. The parametrization comprises the leukemia’s proliferation ($p_i$) and self-renewal ($a_i$) rates, as well as the effectiveness of the treatment components ($k_1, k_2$) on a molecular level. We present two personalization approaches: (1) in the leukemia parameters ($p_i, a_i$) and (2) additionally also in chemotherapy effectiveness ($p_i, a_i, k_1, k_2$). The ODE model and parameter fitting was implemented in R using the deSolve and DEoptim packages. The combination of model-fitted disease characteristics and clinical (study) data allows to investigate AML in more detail. Overall, our approach contributes to a more personalized analysis strategy in clinical research and opens the path to a model-supported decision making in clinical practice.

References

Authors/Institutions: D. Görlich, A. Faldum, Institute of Biostatistics and Clinical Research, University of Münster, Münster, NRW, GERMANY|J.C. Banck, Department of Internal Medicine III, University Hospital Munich, Munich, GERMANY|J. Braess, Department of Oncology and Hematology, Hospital Barmherzige Brüder, Regensburg, GERMANY|
Abstract Body: Large amount of the research in survey sampling has been directed towards improving estimates using information available for the entire population. The efficiency of such methods depends on several factors such as the information available for the entire population, the sampling strategy, the sample size, etc. There is no an absolutely optimal design, but under certain principles and restrictions, a well designed sampling strategy can be implemented. In two-phase designs, efficiency is gained with stratification by auxiliary information known early in the design. There is a number of reasons why investigators may want to stratify: it offers gains in efficiency when the target variable behaves differently between strata and estimates can be obtained for each strata. An further way to gain efficiency is by optimally allocating the resources. For example, conditional on a given sample size or a given precision, what is the optimal allocation of sample sizes? Large amount of the research on optimal allocation has been directed towards estimation of totals or functions of totals. However, in many instances inference is concerned with regression parameters from data that arises from a correlated setting. We examine the impact of ignoring the correlation in when allocating thee resources. Using theory from sampling survey, we extend and propose different allocation methods that allow for correlated data.

AUTHORS/INSTITUTIONS: C. rivera-rodriguez, Statistics, University of Auckland, Auckland, Auckland, NEW ZEALAND}
Abstract Body: Background: Cervical cancer remains the second most commonly diagnosed cancer and the third leading cause of cancer death in developing countries. Improving clinicians' knowledge and understanding of surgico-pathological staging is critical in the fight against the disease. However, a systematic evaluation of different ordinal regression models based on diverse predicted outcomes has not been given its due share in literature.

Objective: To systematically assess the flexibility of odds ratios for three popular ordinal regression models i.e., the Cumulative Proportional Odds model (CPOM), The Continuation Ratio model (CRM) and Adjacent Category Logistic Model (ACLM) when applying cervical cancer data in surgical stage prediction.

Method: We systematically compared the performance of CRM, CPOM and the ACLM as the predictive mechanisms, and evaluate the most appropriate model in the cervical cancer setting. The study considered women who visited the Oncology department at the Moi Teaching and Referral Hospital's Chandaria Cancer and Chronic Diseases Center and were diagnosed and surgically treated for cervical cancer from January 2014 to December 2018.

Results and conclusion: We presented the comparison between 3 different regression models for ordinal data with respect to goodness-of-fit under cervical cancer setting. We found that the CRM model without proportional odds yielded better results. All the 5 independent features selected for classifying the patients into surgical stages had significant adjusted odds ratio with the following predictor variables the FIGO clinical stage, the presence or absence of cancer in the vaginal and parametrial regions, the presence or absence of symptomatic vaginal discharge and lower abdominal pain.

Authors/Institutions: C.O. Odhiambo, SIMS, Strathmore University, Nairobi, KENYA
Abstract Body: In prognosis studies to evaluate association between a continuous biomarker and a survival outcome, investigators often classify subjects into subclasses of high/low-expression groups and applied simple survival analysis techniques of Kaplan-Meier method and the logrank test. The high/low-expressions are defined according to whether the observation of the biomarker is higher than the cut-off value, which is heterogeneous across studies. The heterogeneous definitions of the cut-off makes it difficult to apply the standard meta-analysis techniques. In this talk, we introduce statistical methods to estimate the time-dependent receiver operating characteristic curve and the concordance index for a survival outcome synthesizing published prognosis studies, in which Kaplan-Meier estimates for the high/low-expression groups are reported. We illustrate our proposed methods with a real dataset of meta-analysis of prognosis studies for Ki-67 and evaluate their performance with simulation studies. This is a joint work with Professor Xiao-Hua Zhou.
This talk presents a new goodness-of-fit test for a stereotype model used for an ordinal response variable. The proposed test is based on the well-known Hosmer-Lemeshow test and its version of the proportional odds regression model. This latter test is calculated from a grouping scheme assuming that the levels of the ordinal response are equally spaced, which might be not true. One of the main advantages of the stereotype model is that it allows us to determine a new uneven spacing of the ordinal response categories, dictated by the data. The proposed test takes the use of this new adjusted spacing to partition data. We also show a modification of the proposed test for handling large samples. A simulation study shows good performance of the proposed tests under a variety of scenarios. The results of a health application are presented.

I. Liu, School of Mathematics and Statistics, Victoria University of Wellington, Wellington, Wellington, NEW ZEALAND
D. Fernández, Institut de Recerca, Parc Sanitari Sant Joan de Déu, Barcelona, Barcelona, SPAIN
TITLE: The R package “ClustOrd”: Model-based clustering for ordinal data based on finite mixture models

ABSTRACT BODY:

Abstract Body: Many of the methods that deal with clustering in matrices of data are based on mathematical techniques such as distance-based algorithms or matrix decomposition. In general, it is not possible to use statistical inferences or select the appropriateness of a model via information criteria with these techniques because there is no underlying probability model. Additionally, the use of ordinal data is very common (e.g. Likert, pain, or Braun-Blanquet scale). Recent research has developed a set of mixture-based models for a data matrix of ordinal data. Those approaches apply likelihood-based clustering via finite mixtures to the proportional odds and ordered stereotype models. We will show a recently developed R package called “ClustOrd” which allows users to perform clustering of ordinal data using those models. Fuzzy allocation of rows, columns, and rows and columns simultaneously (biclustering) to corresponding clusters is obtained by performing EM algorithm or Reversible-Jump MCMC. Applications in health will be shown to illustrate the functioning of this R package.

AUTHORS/INSTITUTIONS: L.F. McMillan, School of Mathematics and Statistics, Victoria University of Wellington, Wellington, NEW ZEALAND| D. Fernández, Institut de Recerca. CIBERSAM, Parc Sanitari Sant Joan de Déu, Barcelona, SPAIN|
Abstract Body: Motivated by a dataset on glaucoma progression, we propose a Bayesian Nonparametric model for multivariate longitudinal measurements. A key motivation of the research relates to the assessment of glaucoma progression and the understanding of which factors are more determinant for detecting progression. The model features a vector autoregressive (VAR) model for multivariate profiles, and a logit stick-breaking prior to flexibly incorporate baseline information for the available patients. Extensive simulation studies suggest the utility of the model to detect those more influential predictors for early or late progression. Results for the motivating study are discussed.

AUTHORS/INSTITUTIONS: F.A. Quintana, Statistics, Pontificia Universidad Católica de Chile, Santiago, RM, CHILE|M. Beraha, A. Guglielmi, Mathematics, Politecnico di Milano, Milano, ITALY|
ABSTRACT BODY:

Abstract Body: One problem of classification algorithms are borderline samples with similar scores for different categories. Particularly in the case of classification-based diagnosis, a small change in the score can have significant consequences for further treatment. Our proposal is to deliberately introduce a new category in disease classification — the Undecidable category. A patient falls into the category Undecidable if the available information is insufficient for making a diagnosis that rules out negative consequences for the patient in case of a misdiagnosis. We propose a model based on the Dempster-Shafer Theory of Evidence (DST), a statistical framework developed in the 1960s.

The idea of DST is the extension of the concept of Probability by definition of Belief and Plausibility. These two terms represent a lower and an upper bound of Probability and are collectively referred to as an Evidence. Each event is no longer assigned a one-dimensional Probability but a two-dimensional Evidence. For any such Evidence, the Belief is always less than or equal to the Plausibility and the difference is referred to as Uncertainty.

Our approach is to match Evidences of possible diagnoses against each other. A diagnosis is considered as assured if its Belief is greater than the Plausibilities of all competing diagnoses. The more information that flows into the model, the smaller the Uncertainty of an event becomes, and the more likely it is to meet the conditions for risk-free diagnoses.

While classic probabilistic calculus is predicated on probabilities, the DST offers some very useful degrees of freedom. Thus, there are several possibilities of additive (associative and commutative) evidence combination rules (ECR). We take advantage of this to introduce a comprehensive model to combine evidences based on receptor genes, co-expressed genes, immunohistological measurements and well-known pathways, which also takes into account the importance of these factors. A multiplicative ECR will then link the status of single receptors to their associated breast cancer subtypes. As an additional feature, we allow physicians to involve their opinion by another evidence we call Expert Knowledge.

We treat in detail different strategies for training the required mass functions used by DST. In particular we address the problem of a missing golden standard by using the concept of Uncertainty as defined in DST.

AUTHORS/INSTITUTIONS: M. Kenn, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, AUSTRIA
Abstract Body: Electronic Health Records (EHR) data provide a unique opportunity to understand and assess the health of individuals. While EHR data contain granular information on individuals’ clinical status, the data often lack information on social factors such as one’s environmental context. However, we can supplement EHR data by linking to external data. One readily available data source is environmental data - both geospatial data on the built environment and temporal data on the climate environment. In this talk we describe a study where we used EHR data to identify a cohort of children with asthma (n = 5356). We then geocoded the children to contextualize their built environment (distance to highway, distance to parks, walkability etc.). Additionally, we characterized temporal variability in the form of weather, pollen count, air quality etc. Our goal is to assess how spatial-temporal factors interact to impact risk of asthma exacerbations. We also sought to predict daily risk of asthma exacerbation to provide personalized alerts. Methodologically, we had to resolve issues of clinical phenotyping, data linkage, and analysis of data on different time and location scales. Overall we illustrate how EHR data, linked with publicly available data can provide valuable insights into children’s health.

Authors/Institutions: B. Goldstein, Biostatistics & Bioinformatics, Duke University, Durham, North Carolina, UNITED STATES|B. Goldstein, J. Lang, Children’s Health & Discovery Initiative, Duke University, Durham, North Carolina, UNITED STATES|B. Goldstein, J. Lang, Department of Pediatrics, Duke University, Durham, North Carolina, UNITED STATES
Abstract Body: In the presence of an intermediate clinical event, the analysis of time-to-event survival data by conventional approaches, such as the log-rank test, can result in biased results due to the length-biased characteristics. In the present study, we extend the studies of Finkelstein and Nam & Zelen to propose new methods for handling interval-censored data with an intermediate clinical event using multiple imputation. The proposed methods consider two types of weights in multiple imputation: 1) uniform weight and 2) the weighted weight methods. Extensive simulation studies were performed to compare the proposed tests with existing methods regarding type I error and power. Our simulation results demonstrate that for all scenarios, our proposed methods exhibit a superior performance compared with the stratified log-rank and the log-rank tests. Data from a randomized clinical study to test the efficacy of sorafenib/sunitinib vs. sunitinib/sorafenib to treat metastatic renal cell carcinoma were analyzed under the proposed methods to illustrate their performance on real data. In the absence of intensive iterations, our proposed methods show a superior performance compared with the stratified log-rank and the log-rank test regarding type I error and power.

Authors/Institutions: J. Kim, Applied Statistics, University of Suwon, Hwaseong, Gyeonggi, KOREA (THE REPUBLIC OF)|C. Nam, Preventive Medicine, Yonsei University College of Medicine, Seoul, KOREA (THE REPUBLIC OF)|S. Kim, Clinical Statistics, Yuhan Corporation, Seoul, KOREA (THE REPUBLIC OF)
Many prediction models perform worse when applied to new individuals, which may be caused by the use of invalid model parameters or by differences in case-mix distributions. Because it is increasingly common to develop and validate prediction models using data from different studies, settings and populations, the validity of model predictions may be affected by variation in the samples’ representativeness of the targeted population.

We present propensity-based standardized measures of discrimination and calibration performance, and discuss how these can be used during external validation to assess model transportability and decide upon updating strategies. Further, when developing a new prediction model, we discuss how standardization of available samples improves the estimation of model parameters and subsequent performance across the included settings and populations.

We evaluate the proposed propensity-based standardization methods in an extensive simulation study where we explore under what circumstances standardization of multiple development samples improves prediction model performance. We also explore to what extent it remains possible to assess model performance in particular settings and populations if all data are used during model development. Results demonstrate that combining and standardizing all available samples for development purposes yields more favorable c-statistics and Brier scores even when some of the included studies have case-mix distributions or predictor-outcome associations that do not properly reflect the target population. When no samples are reserved for external validation, performance assessment requires standardized bootstrap procedures to avoid bias and over-optimism. Finally, we illustrate our methods in a motivating example where data from 13 studies were used to develop and externally validate a prediction model to diagnose deep vein thrombosis in patients suspected of deep vein thrombosis.

In conclusion, propensity score-based standardization might help (i) improve the interpretation of external validation studies of existing prediction models, and (ii) enhance the reproducibility of newly derived prediction models across different settings and populations.

**AUTHORS/INSTITUTIONS:** V. de Jong, J. Hoogland, T. Debray, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, NETHERLANDS|L. Nguyen, Department of Public Health, University of Copenhagen, Copenhagen, DENMARK
Proportional hazard Cox regression models are overwhelming used in the analysis of biomedical data when the outcome is a time-dependent event. Physicians interpret research results in terms of hazard ratios (HR). However, the HR interpretation strongly depends on the validity of the underlying survival model. The so-called non-collapsibility implies that HR changes when some relevant variable affecting the event of interest is not included in the model even if both the studied and the non-included variables are independent. In this work, we introduce the generalized hazard ratio (gHR), an alternative index which, for a fixed studied characteristic, represents the excess of the risk of the population with some value on this characteristic respect the population in which this value is one unit less. The new measure has not individual but populational interpretation and it is not affected for non-included variables independent with the one studied. We explore its interpretation in both risk and probability terms and its links with marginal Cox regression models. We also propose a non-parametric estimator for the standard right-censored context.
Abstract Body: There is a long debate in experimental design between the classic randomization design of Fisher, Yates, Kempthorne, Cochran and those who advocate deterministic assignments based on notions of optimality. In non-sequential trials comparing treatment and control, covariate measurements for each subject are known in advance, and subjects can be divided into two groups based on a criterion of imbalance. With the advent of modern computing, this partition can be made nearly perfectly balanced via numerical optimization, but these allocations are far from random. These perfect allocations may endanger estimation relative to classic randomization because unseen subject-specific characteristics can be highly imbalanced. To demonstrate this, we consider different performance criterions such as Efron's worst-case analysis and our original tail criterion of mean squared error. Under our tail criterion for the differences-in-mean estimator, we prove asymptotically that the optimal design must be more random than perfect balance but is not completely random. This result vindicates restricted designs that are used regularly such as blocking and rerandomization. For a covariate-adjusted estimator, balancing offers less rewards and it seems good performance is achievable with complete randomization. Further work will provide a procedure to find the explicit optimal design in different scenarios in practice.

Authors/Institutions: A. Kapelner, Mathematics, Queens College, CUNY, Queens, New York, UNITED STATES|A. Krieger, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES|M. Sklar, Stanford University, Stanford, California, UNITED STATES|U. Shalit, D. Azriel, Technion, Haifa, ISRAEL
The large number of multiple tests in imaging analysis can result in large Type II Error rates that preclude important findings when stringent Type I Error controls are used. Commonly used multiple comparison corrections include family-wise error rate control (e.g., Random Field Theory and Bonferroni correction) and false discovery rate control, both of which depend on p-values. However, stringent reliance on p-values has been criticized recently in the statistics and neuroscience community. A newly proposed method, the Second-generation p-value (SGPV), overcomes interpretability issues with traditional p-values and has good performance characteristics. SGPVs can be interpreted directly as the proportion of estimates supporting the null hypothesis. Further, by specifying a clinically meaningful region beforehand, the Type I Error rate is naturally controlled and encompasses a proper scientific correction for multiple comparison. Lastly, the pre-specified clinically meaningful region promotes good research practice and prevents the post-hoc interpretation of mediocre results.

For functional imaging analysis, we construct the null interval with the data observed in functionally null region, the cerebrospinal fluid (CSF). In this study, we evaluate the usage of SGPVs in group inference compared to other traditional multiple correction methods. An R shiny app is developed for general usage and easy visualization.
Abstract Body: Linear mixed models (LMMs) are widely used to describe the relationship between a response and explanatory variables when the data are clustered. The effect of the cluster structure in the response variable is incorporated by including random effects in the model; cluster structure in the explanatory variables is often ignored, possibly because interpreting the model conditionally makes it seem unnecessary to consider this structure. We study in detail the effects of fitting the standard two-level LMM with a single explanatory variable to clustered data. This model ignores clustering in the explanatory variable and we explore the effect of (the usually ignored) within-cluster correlation in the explanatory variable.

In this talk, we present a number of unexpected findings. (1) Ignoring clustering in the explanatory variable affects estimators of both the regression and variance parameters and the effects are different for different estimators. (2) Increasing the within cluster correlation of the explanatory variable introduces a second local maximum into the log-likelihood and reduced or restricted maximum likelihood (REML) criterion functions which eventually becomes the global maximum, producing a jump discontinuity (at different values) in the maximum likelihood (ML) and REML estimators of the parameters. (3) Standard statistical software can return local rather than global ML and REML estimates in this very simple problem. (4) Local ML and REML estimators may fit the data better than their global counterparts but, in these situations, ordinary least squares (OLS) may perform even better than the local estimators. We also establish central limit theorems hold for the ML and REML estimators of the parameters in misspecified linear mixed models which are of some independent interest.

Authors/Institutions: A. Welsh, H. Yoon, The Australian National University, Canberra, Australian Capital Territory, AUSTRALIA
Abstract Body: Analysts of clustered and longitudinal data find conditional likelihood methods and closely related generalized linear mixed models that decompose covariates into between- and within-cluster components appealing not only for their useful statistical properties but because they provide covariate effect estimates of scientific interest. However, missing data and drop out are common in longitudinal studies and predictors can be correlated with subject-specific effects. Statistical methods that fail to accommodate these correlations, such as standard mixed effects models, lead to inconsistent estimators. Other approaches, such as conditional likelihood, are consistent when there are correlations between predictors and subject-specific effects but are inconsistent with data that are missing at random. This inconsistency can lead to incorrect scientific conclusions. There is confusion in the literature about the performance of standard statistical methods in these settings.

In this talk we clearly identify settings where conditional likelihood and covariate decomposition methods provide consistent estimation of covariate effects of interest. Theory, simulation studies and results of the analysis of data from a longitudinal study of physical and cognitive functioning illustrate the findings. Our work provides easy-to-use strategies to avoid bias and misleading findings.

Authors/Institutions: J. Neuhaus, C. McCulloch, Division of Biostatistics, University of California, San Francisco, San Francisco, California, UNITED STATES
Abstract Body: Longitudinal data are sometimes collected over long time periods to examine associations between risk factors and outcomes, as associating a risk factor change within a person with change in outcome is considered more indicative of a causal relationship than comparisons between individuals with and without a risk factor. The latter is often feared to be confounded by contextual effects. However, all longitudinal modeling techniques are not equally capable of adjusting for contextual effects. In fact, most biostatisticians fit overall effects via either random effects models or GEE that combine between and within comparisons in design and modeling dependent ways.

Random effects models can easily be spliced into orthogonal within and between individual effects and provide significance tests of whether these differ. However, subject matter knowledge and further investigation is needed to determine whether such differences are necessarily indicative of contextual effects. Non-linearity in continuous covariates, period effects, carry over effects or changes in measurement technique and measurement error in covariates can cause similar differences. In fact, if it is determined that the effect difference is due to measurement error, this can be used to correct the association for this error.

In this talk, I cover some background as well as models for how effect discrepancies arise and the magnitude of biases when effects are combined in a standard black box manner and also show examples from several longitudinal studies.

AUTHORS/INSTITUTIONS: M. Palta, Department of Population Health Sciences/ Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, UNITED STATES
Hierarchically structured data arises frequently in practice, in many fields of science and social science. Linear mixed models (LMMs) is one of the most common statistical methods to deal with the structured or clustered data. Yoon and Welsh (2019) studied the effect of ignoring clustering in x on fitting LMM, and showed that it can be obtained misleading assessments of both the association between y and x and of the variance components and showed that, as the within cluster variance of x, \( \tau_x \), increases, the likelihood and the REML criterion develop two distinct local maxima and which of these is the global maximum changes at the jump point. In LMMs for clustered or hierarchical structured data, regressors can be correlated with random effects. When the random effects and regressors independence assumptions are violated, not only regression coefficient estimators but also variance components can be severely biased. In this study, we investigate how the violation of the random effects and regressors independence assumption could affect on the estimates of parameters and compare the results with the effect of ignoring clustering in x when fitting LMM (Yoon & Welsh).

**AUTHORS/INSTITUTIONS:**  H. Yoon, The Australian National University, Canberra, Australian Capital Territory, AUSTRALIA|Y. Lee, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|A. Welsh, Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australian Capital Territory, AUSTRALIA|
Master protocol designs are often proposed to improve the efficiency of drug development with multiple subgroups. In the basket trial design, different subgroups can have similar biological pathogenesis pathways. Hence, a target therapy can result in similar responses. A good information sharing strategy between different subgroups can potentially improve the efficiency of evaluating treatment efficacy. In traditional hierarchical models, all subgroups are placed into the same sharing pool for information sharing under the exchangeability assumption. However, due to the heterogeneity between subgroups, there can be large differences in drug efficacy. Under such cases, strong borrowing across all subgroups is not suitable and no borrowing can be inefficient, because the treatment effect is analyzed in each subgroup separately. We propose a Bayesian Cluster Hierarchical Model (BCHM) to improve the operating characteristics of estimating the treatment effect in multiple subgroups in basket trials. Bayesian nonparametric method is applied to dynamically calculate the number of clusters by conducting a multiple cluster classification based on subgroup outcomes. A hierarchical model is used to compute the posterior probability of the treatment effect, with the borrowing strength determined by the Bayesian nonparametric clustering and the similarities between subgroups. For treatment effect estimation, BCHM provides lower MSE values compared to traditional hierarchical models. We apply BCHM and construct a loss function to optimize the design parameters. BCHM provides a balanced approach and smart borrowing, which yields better results in assessing the treatment effect in different scenarios compared to other conventional methods.

AUTHORS/INSTITUTIONS: J. Lee, N. Chen, Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, UNITED STATES
Abstract Body: Many prognostic models are created using survival data. In practice, the development of such models remains fairly ad hoc, and the temporal aspect of survival data is often underused. I will outline a number of existing methods for evaluating prognostic survival models. In particular, the emphasis will be on tools that can quantify how prognostic performance varies with time. I will also present a complementary new tool we have developed, the hazard discrimination summary (HDS; Biometrics 2017). HDS is an interpretable, risk-based measure of how a model’s discrimination varies with time. I will also describe a connection between HDS and the Cox model partial likelihood.

Authors/Institutions: C.J. Liang, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Rockville, Maryland, UNITED STATES|P. Heagerty, Department of Biostatistics, University of Washington, Seattle, Washington, UNITED STATES|
Almost 100 years ago, Sewell Wright gave the connection between allelic correlation and identity by descent (IBD) at a single locus in an idealized population. For many years the data did not allow for direct studies, and the only measures of IBD were expectations based on known pedigree structures, but it was already clear that the connection between genotypic correlation and IBD is weak when considering pairs of individuals beyond close family relationships.

The ease of computing pairwise genotypic correlations with modern SNP data, and the clarity of the basic formulae of Wright have led to wide use of the matrix of realized pairwise genotypic correlations (the GRM) being used as a measure of pairwise realized “relatedness” in large population samples from which apparent close relatives are often removed. Many forms of the GRM, involving weighting or other possible kernel-based adjustments, have been successfully used in studies of quantitative variation and in phenotypic prediction. As such these approaches are best viewed as regression approaches with random or fixed effects, and a very large number of predictors, rather than as deriving from ancestral identity by descent. The GRM can also provide measures of relatedness and structure among populations with divergent allele frequencies. We show that covariances of genotypes provide good measures of population-level relatedness if multiple samples are available from populations that have diverged due to random genetic drift.

However, the GRM is a poor estimate of realized IBD between pairs of individuals, and realized genome-wide IBD varies substantially about its pedigree expectation. We show that attempts to connect the GRM to individual relatedness and particularly to remote pedigree relationships are misguided. On the other hand SNP data provide location-specific information on segments of genome shared by descent. In remote relatives, segments of IBD are rare but not short. Although many relatives will share no IBD DNA, those segments of IBD that exist can be detected from SNP data. We show that location-specific inference of IBD segments jointly among relatives can provide information to resolve genetic traits, to detect regions of the genome subject to selection, and to model processes that go beyond a simple additive model.

**AUTHORS/INSTITUTIONS:** E.A. Thompson, Department of Statistics, University of Washington, Seattle, Washington, UNITED STATES|
While machine learning methods may have the potential for higher accuracy in predictions than classical statistical methods due to their flexibility, their results usually cannot be presented by a transparent formula. Hence, it is often not clear how specific values of the predictors lead to the predictions. We set up a competition between machine learning and classical statistical methods to model risk for cardiovascular events within one year from a health screening. In this competition, we compared these methods in terms of interpretability, variability of predictions, and prediction performance varying the size of the training set from small to very large.

The training set comprised well established risk factors for cardiovascular disease and additional potential risk factors from one million participants of the Austrian health screening program. We applied different modelling approaches to this training set: a logistic regression model with splines and interactions using R, neuronal networks using Matlab, and extreme gradient boosting with regression trees using Python. Subsequently, we divided the training set into subsets (2, 10, 25 and 100 random splits), adapted the modelling strategy of each approach to the sample size, and repeated the adapted modelling procedure in each subset. Using individual conditional expectation and partial dependence plots, we assessed the functional forms of the association of prognostic factors with predictions and their variability across the subsets. Model performance (c-index and calibration) was assessed in a test set comprising 500,000 individuals.

We found that functional forms were relatively similar across the modelling approaches. Extreme gradient boosting using regression trees showed the wiggliest functional forms that could not be rationally explained. The variability of predictions was higher for machine learning methods than for the logistic regression model when the sample size was lowered, pointing to overfitting. However, with large samples machine learning methods slightly outperformed logistic regression in terms of model test performance.

In a setting with large samples and few potential predictors, carefully parameterized machine learning and flexible statistical modelling lead to similar conclusions on the role of predictors. In small samples, however, naively applied machine learning can lead to overfit.

**AUTHORS/INSTITUTIONS:** C. Wallisch, A. Agibetov, G. Dorffner, D. Dunkler, G. Heinze, Medical University of Vienna, Vienna, AUSTRIA; C. Wallisch, Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Berlin, GERMANY
Abstract Body: We often face a huge challenge in properly analyzing aggregate ordinal outcome data when there is no information to separate certain two subcategories exactly. In this talk, we propose a network meta-regression approach for modeling ordinal outcomes with uncertain categories in order to evaluate efficacy of treatments for Crohn’s Disease. Specifically, we develop regression models based on aggregate trial-level covariates for the underlying cut points of the ordinal outcomes as well as for the variances of the random effects to capture population heterogeneity across trials. Our proposed models are particularly useful for indirect comparisons of multiple treatments that have not compared head-to-head within the network meta-analysis framework. Moreover, we introduce Pearson residuals and construct an invariant test statistic to evaluate goodness-of-fit in the setting of ordinal outcome meta-data. A case study demonstrating the usefulness of the proposed methodology is carried out with creating three different allocations of counts in those uncertain categories. We also conduct a full Bayesian analysis via the latent count approach.

Authors/Institutions: Y. Gwon, Biostatistics, University of Nebraska Medical Center, Omaha, Nebraska, UNITED STATES|J. Ibrahim, Biostatistics, University of North Carolina, Chapel Hill, North Carolina, UNITED STATES|J. Li, Eli Lily, San Diego, California, UNITED STATES|M. Chen, Z. Chi, Statistics, University of Connecticut, Storrs, Connecticut, UNITED STATES|M. Mo, A. Xia, Amgen Inc, Thousand Oaks, California, UNITED STATES|
For quantitative traits of agronomic importance influenced by a large number of genes with small effects, genomic-assisted selection has been an invaluable tool for understanding the relative genetic performance of varieties/lines in breeding populations. One of the challenges of genomic-assisted prediction with large-p-with-small-n regressions is the development of robust and efficient approaches that accurately predict phenotypic traits as functions of genotypic and environmental inputs. With advancements in phenotyping and genotyping technologies, multi-trait multi-environment data are now commonly generated in plant breeding programs. However, for this type of data, there is a distinct lack of statistical modelling approaches that can exploit the phenotypic and genetic dependencies and improve the accuracy in the context of genomic selection. In the present study, we propose the multi-environment models for genomic prediction with the integration of genome wide association analysis and genomic best linear unbiased prediction using genomic data and multi-trait multi-environment phenotype data in barley.

AUTHORS/INSTITUTIONS: K. Chen, SAGI West, School of Molecular and Life Sciences, Curtin University, Bentley, Western Australia, AUSTRALIA|C. Hill, T. Angessa, X. Zhang, C. Li, Western Barley Genetics Alliance, School of Veterinary and Life Sciences, Murdoch University, Murdoch, Western Australia, AUSTRALIA|
Analysis of multivariate longitudinal immuno-epidemiological data using a pairwise joint modelling approach

Abstract Body: Background: Immuno-epidemiologists are often faced with multivariate outcomes, measured repeatedly over time. Such data are characterised by complex inter- and intra-outcome relationships which should be accounted for during analysis. Scientific questions of interest might include determining the joint effect of a treatment on the evolution of all outcomes together, or grouping outcomes that change in the same way. Modelling different outcomes separately may not be appropriate because it ignores the underlying relationships between them. In such situations, a joint modelling strategy is necessary. This work describes a pairwise modelling approach and discusses its benefits over more simple analytic approaches, with application to data from a study of the response to BCG vaccination in the first year of life, conducted in Entebbe, Uganda.

Methods: The study aimed to determine the effect of maternal latent Mycobacterium tuberculosis infection (LTBI) on infant immune response (TNF, IFN-$\gamma$, IL-13, IL-10, IL-5, IL-17A and IL-2 responses to PPD), following immunisation with BCG. A simple analysis ignoring the correlation structure of multivariate longitudinal data is first shown. Univariate linear mixed models are then used to describe longitudinal profiles of each outcome, and are then combined into a multivariate mixed model, specifying a joint distribution for the random effects to account for correlations between multiple outcomes. Due to model fitting complexity because of the high number of outcomes, a pairwise joint modelling approach, where all possible pairs of bivariate mixed models are fitted, is used to obtain parameter estimates.

Results: Both univariate and pairwise longitudinal analysis approaches are consistent in finding that LTBI had no impact on the evolution of cytokine responses to PPD. Estimates from the pairwise joint modelling approach were more precise. Advantages of the pairwise approach include the opportunity to test for the effect of LTBI on the joint evolution of all outcomes and the ability to estimate association structures of the outcomes.

Conclusions: The pairwise joint modelling approach reduces the complexity of analysis of high-dimensional multivariate repeated measures, allows for proper accounting for association structures and can improve our understanding and interpretation of longitudinal immuno-epidemiological data.

Authors/Institutions: L. Lubyayi, J. Levin, Department of Epidemiology and Biostatistics, School of Public Health, University of the Witwatersrand, Johannesburg, SOUTH AFRICA|L. Lubyayi, P.A. Mawa, S. Cose, A.M. Elliott, Immunomodulation and Vaccines Programme, Medical Research Council / Uganda Virus Research Institute and London School of Hygiene and Tropical Medicine Uganda Research Unit, Entebbe, UGANDA|P.A. Mawa, Uganda Virus Research Institute, Entebbe, UGANDA|S. Cose, A.M. Elliott, Department of Clinical Research, London School of Hygiene and Tropical Medicine, London, UNITED KINGDOM|E.L. Webb, MRC Tropical Epidemiology Group, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UNITED KINGDOM|
The standard survival model assumes that, if there is no censoring, at a future time point, all individuals will experience the event of interest. However, cure models have been developed because there might be situations where this assumption is not appropriate. For example, in clinical settings, it is very unlikely to have any recurrence of some tumors later than a certain period after radiation treatment. In sociology, there are many examples of this kind of events: marriage, birth of second child, career shift, etc. In most literature, the subjects whose events will not occur are referred to as cured subjects.

Mixture cure models assume that the population is a mixture of cured and susceptible individuals. The common assumption of traditional cure models is that cured and uncured subjects cannot be distinguished within the censored observations. Hence, the cure indicator is usually modelled as a latent variable. However, there might be situations when extra information about cure status is available. For instance, an individual is assumed to be cured or a long-term survivor if the observed survival time is greater than the cure threshold or based on diagnostic tests.

This aim of this paper is twofold: first, we propose a novel nonparametric estimator of the conditional survival function in the mixture cure model when cure status is partially known. The second goal is to propose a nonparametric estimator of the cure probability when cure status is partially known. The asymptotic representations are obtained, from which strong consistency and asymptotic normality of the estimators are derived. Using simulation studies, we have shown that, if a bandwidth parameter is suitably chosen, our estimators perform better than others in an ample range of covariate values. For this reason, a bootstrap bandwidth selector is proposed. The practical performance of these estimators is shown using a dataset of cancer patients with sarcoma.
The influence that environments exert on genetic traits of agricultural crops can interfere with productivity. Thus, it is of paramount importance researches aimed at detecting and analyzing the interactions between genotypes and environments (GEI) and between quantitative trait locus (QTL) and environment (QEI) in multi-environmental trials (METs). Two of the most fixed effects models used to model and understand GEI are the additive main effect and multiplicative interaction (AMMI) model and the genotype main effects + genotype environment interaction (GGE) model, which have proven to provide good results. However, these models are not adequate when the phenotypic data shows heterogeneous error variance throughout the environments, which is common in METs. For the AMMI model, previous work was conducted in order to overcome this limitation. In this paper, we present a generalization of the GGE model that accounts for data with heterogeneous error variance, the weighted GGE (W-GGE) model. The results, for both GEI and QEI, obtained with the W-GGE model are compared with the results from the GGE model and from the AMMI, weighted AMMI and linear mixed model. To conduct this comparison, two data sets are used: data from a simulated pepper (Capsicum annuum L.) back cross population using a crop growth model to relate genotypes in a non-linear way to phenotypes, and the doubled haploid Steptoe x Morex barley (Hordeum vulgare L.) population. Considering the number of detected QTLs and the values for their LOD scores the proposed W-GGE model has proved to provide better results than the unweighted competing methods. (Joint work with Tatiana Assis and Carlos Tadeu Dias)

Authors/Institutions: P.C. Rodrigues, Federal University of Bahia, Salvador, Bahia, BRAZIL
In the past years, microbial research has made enormous progress with the advent of sequencing technologies to investigate the roles of all microorganisms in different ecological habitats. However, microbiome studies (based on 16S amplicon and shotgun sequencing) are difficult to replicate as they may suffer from different sources of batch effects. In this context, we define batch effect as any unwanted source of variation that is unrelated to, but obscures the biological factor of interest. Such batch effects range from biological, technical to computational factors. Batch effect correction is challenging in microbiome data because of the inherent characteristics of data, including sparsity, overdispersion, uneven library sizes, correlation between variables, compositional nature and small sample size. Thus, traditional statistical methods developed for microarray or RNA-seq data are not suitable in this context.

We propose two novel computational methods PLSDA-batch and sPLSDA-batch based on Partial Least Squares Discriminant Analysis regression to address some of these challenges. Data are transformed with centered log ratio to account for uneven library sizes and compositional constraints. We estimate latent components associated first with the outcome, then with batch effects. Using deflation, the batch effect is then subtracted from the original data while the effects of interest remain. The variant sPLSDA-batch includes lasso penalisation to select relevant microbial features. Both methods are non-parametric and can handle the skewed distribution caused by sparsity and overdispersion. Their multivariate property accounts for the data correlation structure.

On both simulated and real microbiome data, we showed that our approaches are superior in removing batch variation compared with existing methods such as ComBat, removeBatchEffect and Remove Unwanted Variation III (from the sva, limma, ruv packages). PLSDA-batch preserves more variation associated with the outcome, while sPLSDA-batch removes batch effects and selects relevant discriminative microbial variables.

Our proposed methods are useful alternatives to address batch effects in microbiome data and will be available through an R package.

Y. Wang, K. Lê Cao, Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, AUSTRALIA
CONTROL ID:  3341110
TITLE:  Modeling international migration flows by integrating multiple data sources
ABSTRACT BODY:
Abstract Body:  Migration has become a significant source of population change at the global level, with broad societal implications. Although understanding the drivers of international mobility is critical to enacting effective policies, theoretical advances in the study of migration processes have been limited by the lack of data on flows of migrants, or by the fragmented nature of these flows. Specifically, data may suffer problems of reliability, which are due to several temporal definitions of migration events, issues of under-registration and imperfect population coverage, and varying accuracy of the data collection systems. Moreover, there are problems of incompleteness, from missing data for specific flows in specific years to countries completely lacking data for one particular data source. Hence, we build on existing Bayesian modeling strategies to develop a statistical framework for integrating different types of data on migration flows. We offer estimates, as well as associated measures of uncertainty, for immigration, emigration, and net migration flows among 31 European countries, by combining administrative and household survey data from 2002 to 2015.

Our framework is made up of two parts. First, we develop a measurement error model to account for the inconsistencies in data measurement with specific parameters. Second, we tackle the issue of data incompleteness by using a migration model based on both migration theory and empirical evidence to estimate the missing data. Such a model employs auxiliary data for pairs of countries informing on the relative attractiveness of each country as a destination. Finally, we include assumptions and expert judgment regarding data quality in the forms of prior distributions for the parameters to supplement the incomplete migration data.

Substantively, we document the historical impact of the EU enlargement and the free movement of workers in Europe on migration flows. Methodologically, our approach improves on previous work on the integrated modeling of European migration data, where data usually comes from a single source. We believe that our statistical framework for evaluating recent migration trends is robust and flexible enough to be further extended to incorporate new data sources, like social media.

AUTHORS/INSTITUTIONS:  E. Del Fava, E. Zagheni, Max Planck Institute For Demographic Research, Rostock, GERMANY|A. Wisniowski, School of Social Sciences, University of Manchester, Manchester, UNITED KINGDOM|
Incorporating a large number of predictors in statistical models introduces issues with model fitting including variable selection, feature extraction and multi-collinearity. Inferential challenges can also result, with the model fitting process needing to address overfitting and false discovery rates. Additionally, when data arise from designed experiments, variance partitioning associated with the design strata defined by the randomisation process needs to be accommodated.

In the area of field crops research, understanding the drivers of the variety by environment (VxE) interaction through the use of environmental covariates is highly desirable. Moreover, the potential to predict variety performance in an untested environment is invaluable to the future of agronomic research.

Previously, a number of statistical methods have been proposed to explain the VxE interaction via environmental covariates. Most of these methods do not account for experimental design terms, spatial field trend and heterogeneity of varietal and residual variance across experiments in the first stage of analysis. Furthermore, current methods are based on a two-stage approach, where uncertainty in estimates from the first stage of analysis are not carried through to the second stage model for the VxE interaction.

A one-stage multi-environment trial analysis approach is presented, including environmental covariates to describe the VxE interaction. The linear mixed model framework allows for heterogeneous residual variance across experiments and adjustments for design effects and spatial field trend. The motivating dataset is composed of 17 environments across the Northern grains region of Australia. Detailed data for 26 environmental covariates representing different weather, soil and crop management conditions was captured within each experiment. Covariate data was obtained at different strata, with some covariates measured on the experimental unit, and others at the variety and experiment level. Covariates were incorporated into the model via forward selection, to identify the most important predictors, avoid multi-collinearity, and ensure a parsimonious model.

This study is the first phase in developing one-stage statistical models that can assist in determining the key drivers of the VxE interaction and can be utilised by crop modellers to refine model predictions.

**AUTHORS/INSTITUTIONS:** M.H. Mumford, K. Bell, C. Forknall, D. Aisthorpe, A. Kelly, Department of Agriculture and Fisheries, Toowoomba, Queensland, AUSTRALIA|D. Rodriguez, J. Eyre, Queensland Alliance for Agriculture and Food Innovation, Gatton, Queensland, AUSTRALIA|L. Serafin, NSW Department of Primary Industries, Tamworth, New South Wales, AUSTRALIA|
The proliferation of electronic data sources, including electronic health records (EHR), has generated new opportunities for using real world data (RWD) to answer novel questions. RWD offer the potential to investigate health-related questions in real-world contexts, improving the efficiency and generalizability of health research. However, RWD are plagued by data quality issues that potentially undermine the validity of research results. EHR data frequently feature high levels of missingness and measurement error due to the complex and messy process that gives rise to these data. Data quality issues for EHR include inconsistent capture of data elements, complex patient-driven missingness mechanisms, and error in exposures, confounders and outcomes. In this talk, we focus on challenges to analysis posed by imperfectly ascertained outcome measures and confounder variables. We demonstrate alternative approaches to accounting for error in EHR-derived variables and compare performance using simulation studies and an analysis of data from a real-world cancer outcomes study. The objective of this presentation is to call attention to data quality challenges in the analysis of RWD, focusing on EHR, and propose alternative approaches that have the potential to mitigate these challenges.
Longitudinal semi-continuous data, characterized by repeated measures of a large portion of zeros and continuous positive values, are frequently encountered in many applications including biomedical, epidemiological, and social science studies. Two-part random effects models (TPREM) have been used to investigate the association between such longitudinal semi-continuous data and covariates. The existing TPREM relies on the form of a (generalized) linear mixed effects model which typically assumes linearity of covariate effects and normality of random effects. These assumptions are, however, limited when the covariate effects are nonlinear and interactive and the between-subject heterogeneity is more complicated than what can be described by a single normal distribution. In this article, we propose a nonparametric Bayesian two-part random effects model (NB-TPREM) to flexibly characterize the association between longitudinal semi-continuous outcome and various covariates. The proposed model relaxes the parametric assumptions to describe complex covariate effects and heterogeneity. Moreover, the NB-TPREM incorporates a time-varying covariate as a functional covariate to examine a longitudinal relationship between the semi-continuous outcome and a time-varying covariate. The methodology is illustrated through a simulation study and an application to social insurance expenditure data collected by the Korean Welfare Panel Study (KOWEPS).
Abstract Body: Recently, technological innovations have produced a lot of complex and high-dimensional data. One example is functional data - in which the unit of observation is a curve or set of curves - and there have been rapid developments in the field of functional data analysis (FDA). There are various topics in FDA and we focus on functional clustering. Rich literature exists on the univariate functional clustering but the methodology for multivariate functional clustering has been less studied. Moreover, no study is available for multivariate functional clustering which incorporates additional covariate information to help identifying the clustering structure correctly. In this research, we propose a Bayesian nonparametric sparse latent factor model for multivariate functional clustering which can optionally incorporate multiple observed covariates. Our model represents the multivariate functional data using a basis expansion and sparse latent factor modeling for the basis coefficients. To induce sparsity, we apply the Indian Buffet Process (IBP) as a prior on factor loadings to induce a spike and slab sparsity. This “spike and slab” sparsity can improve the clustering performance as the factors irrelevant to a particular dimension are not included and more interpretable because a small number of factors are involved. The latent factors and covariates are jointly modeled to follow a Dirichlet process mixture (DPM) of normal distributions, which induce a model-based clustering of the multivariate functional data jointly with the covariates. While our primary focus is on functional clustering, the proposed model can also be viewed as a highly flexible nonparametric functional regression model. From the joint model of the latent factors and covariates, the conditional model can be derived and used to investigate the dependence of response on predictor and make prediction for responses given predictors as functional regression problems. Our method is evaluated and compared with the existing methodologies through various simulation studies and applied to time series data for temperature and precipitation in Canada.
Abstract Body: The ordered stereotype Logit model is a regression model for estimating the effects of predictor variables on ordered categorical responses. Unlike the cumulative model for ordered responses, the stereotype model assumes effects of the covariates which depend on the categories driven by a category specific parameter times a parameter which is invariant over categories. In contrast to the sequential model for ordered responses, the probability of observing a specific category is not modeled via binary models but using a multinomial model formulation. Unlike in the multinomial model, the effects of the covariates are assumed to be proportional and the predictor values are invariant over the categories. Thus, the ordered stereotype logit model can be interpreted as a compromise between the ordered and the unordered categorical logit model. In this talk we present a generalized estimating equations approach to estimate the ordered stereotype panel logit model, where we use a 'working' correlation matrix which is close to the idea in Liang and Zeger (1986). We also propose a corrected variance estimator to overcome problems with biased variance estimation in moderate to small samples if the standard GEE variance estimator is adopted. In an extensive simulation study we a) illustrate the properties of the GEE estimator in various scenarios and b) compare the GEE estimator of the stereotype model using different working correlation matrices. We finally apply the GEE estimator to a data set to estimate the effect of a pharmacological treatment vs. a placebo therapy on rheumatoid arthritis reported on a five-level scale (Lipsitz, Laird and Harrington, 1991).

Authors/Institutions: M. Spiess, Institute of Psychology, University of Hamburg, Hamburg, GERMANY|D. Fernández, Institut de Recerca, Parc Sanitari Sant Joan de Déu, Barcelona, Barcelona, SPAIN|T. Nguyen, I. Liu, School of Mathematics and Statistics, Victoria University of Wellington, Wellington, Wellington, NEW ZEALAND|
Abstract Body: Standard survival models introduce covariates through a single (scale) parameter, and we refer to this standard practice as Single-Parameter Regression (SPR). In contrast Multi-Parameter Regression (MPR) allows covariates to enter the model through multiple distributional parameters, i.e., scale and shape, and Burke and MacKenzie (2017) highlighted its flexibility in the context of survival data. We extend their work to handle multivariate survival data by introducing random effects in both the scale and the shape regression components. We consider a variety of possible dependence structures for these random effects (independent, shared, and correlated), and estimation proceeds using a h-likelihood approach. As the shape parameter may be viewed as a dispersion parameter for log-time, our proposal bears similarities to Double Hierarchical Generalized Linear Modelling (DHGLM). We investigate the performance of our estimation procedure using simulated data, and also consider a real data example.

Authors/Institutions: K. Burke, University of Limerick, Limerick, IRELAND
RNAseq has become the standard technology in gene expression studies in the past few years. It is considered superior to microarrays that used to be the choice of technology in the 2000s. Since RNAseq data are typically summarized as counts per gene for downstream statistical analyses, there have been active developments of statistical models based on negative binomial regression models (NB). To overcome the shortfalls of current NB-based models, we extended the double hierarchical generalized linear models to high dimensional counting data such as RNAseq data and developed an R package for model fitting (DHGLMseq). In addition, we extended Lee and Bjoenstad’s false discovery rate (FDR) control for linear mixed models to the high dimensional DHGMLs. In this presentation, we will review a brief history of advancement of statistical methods for RNAseq data and compare their power and false discovery rates by simulations.
Confidence and likelihood are fundamental statistical concepts with distinct technical interpretation and usage. Confidence is a meaningful concept of uncertainty within the context of confidence-interval procedure, while likelihood has been used predominantly as a tool for statistical modelling and inference given observed data. The likelihood has been extended by accommodating random unknowns via extended likelihood such as h-likelihood. Its methods have been developed for inferences for random unknowns. Here we show that confidence is in fact an extended likelihood, thus giving a much closer correspondence between the two concepts. This result gives the confidence concept an external meaning outside the confidence-interval context, and vice versa, it gives the confidence interpretation to the likelihood. In addition to the obvious interpretation purposes, this connection suggests two-way transfers of technical information. For example, the extended likelihood theory gives a clear way to update or combine confidence information. On the other hand, the confidence connection gives the extended likelihood direct access to the frequentist probability, an objective certification not directly available to the classical likelihood. This implies that intervals derived from the extended likelihood have the same logical status as confidence intervals, thus simplifying the terminology in the inference of random parameters.

AUTHORS/INSTITUTIONS: Y. Lee, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|Y. Pawitan, Karolinska Institutet, Stockholm, SWEDEN
Abstract Body: In resource-limited settings, a cost-efficient approach is to implement a two-phase design. Most statistical methods for the analysis of two-phase designs focus on settings where individual observations are independent. However, it is often the case that study participants are naturally correlated. For instance, patients are treated by clinics and they are correlated within a clinic or observations can be correlated when study participants have repeated measures. Ignoring correlation among observations may underestimate the true variance and may produce test statistics that are wrongly read as significant or not-significant, which results into invalid inferences or conclusions. Therefore, it is important to consider dependencies among the observations for estimation and inference. A recent method for the analysis of marginal models when the data exhibits correlation and arises from a two-phase design is weighted generalized estimating equations (WGEE). In this work, we consider the case of longitudinal data with missing information.

We compare two approaches to estimating the sampling weights: 1. A logistic model; 2. Calibrated weights. We illustrate the methods using longitudinal data from New Zealand on knee replacements performed between 2010 and 2015.

AUTHORS/INSTITUTIONS: Y. Kim, Statistics, The University of Auckland, Auckland CBD, NEW ZEALAND
Abstract Body: There are many situations where one expects an ordering among $K \geq 2$ experimental groups or treatments. Although there is a large body of literature dealing with the analysis under order restrictions, surprisingly, very little work has been done in the context of the design of experiments. Here, a principled approach to the design of experiments with ordered treatments is provided. In particular, we propose two classes of designs which are optimal for testing different types of hypotheses. The theoretical findings are supplemented with thorough numerical experimentation and a concrete data example. It is shown that there is a substantial gain in power, or alternatively a reduction in the required sample size when an experiment is both designed and analyzed using methods that account for order restrictions.

AUTHORS/INSTITUTIONS: O. Davidov, Statistics, University of Haifa, Haifa, ISRAEL
What defines a cell type? This fundamental question in molecular biology remains difficult to answer despite our great advances in sequencing technologies. Single-cell assays are proving to be indispensable to understand cellular heterogeneity, to decompose tissues into cell types and uncover new biology. They have had a broad impact already on fields as diverse as neurobiology or development. However, while these technologies generate large amounts of rich data, our understanding of the unobservable complexity of biological processes is still limited as we are lacking computational and statistical tools to extract relevant information from these data.

Cell identity is defined as a cell’s capacity to perform a unique combination of context-dependent molecular functions. The range of cell’s functional abilities is constrained and regulated by complex regulatory circuits, many of which act through regulation of RNA expression. Single-cell RNA sequencing (scRNA-seq) can be used as a reliable approximation of cell’s unique functional molecular profile and cell’s identity, but the data are sparse, zero-inflated and highly dimensional. Our aim is to identify gene co-expression modules that underpin cell identities across tissues. Different module detection methods can be used to infer gene regulatory networks, for example using clustering to group genes based on global similarity in their expression profiles (co-expression), bi-clustering to group genes and samples simultaneously, network inference to model regulatory relationships between genes. Here we focus on matrix decomposition techniques with independent component analysis, combined with generalised linear models.

We analysed scRNA-seq data from Tabula Muris sequenced with either 10X or SMART-seq2 platforms, each dataset includes ~20,000 cells from 20 different tissues, representing up to 100 different cell types, measured across > 10,000 genes. We focused on specific type of blood cells (macrophages, myeloid, endothelial) across multiple tissues and identified gene modules that characterise these cell types. We compared our approach with alternative existing methods and assessed the predictive ability of our gene modules across different data sets. Ultimately, our approach aims to infer putative transcriptional regulators behind the cell type modules, as we illustrate via biological interpretation.

AUTHORS/INSTITUTIONS: K. Lê Cao, A. Dakic, Mathematics and Statistics, University of Melbourne, Parkville, Victoria, AUSTRALIA
Abstract Body: Longitudinal and high-dimensional measurements have become increasingly common in biomedical research. In survival analysis, longitudinal covariates are usually accounted for using joint models for longitudinal and time-to-event outcomes, while penalized survival models are employed to tackle high-dimensional sets of predictors. However, methods to predict survival outcomes using covariates that are both longitudinal AND high-dimensional are currently missing. In this presentation we propose a strategy to overcome this limitation.

Our work is motivated by data from the MarkMD study, which involved 157 patients affected by Duchenne Muscular Dystrophy (DMD), a rare genetic disorder that leads to loss of ambulation (LoA) and premature death. No cure for DMD is available yet and non-invasive markers are urgently needed to monitor disease progression and predict disease milestones. In the MarkMD study, an antibody-based bead array was used to measure the serum levels of 118 different proteins with 240 antibodies in a total of 303 serum samples. The aim of the study was to understand if longitudinal protein data could be used to improve the prediction of the age at which LoA occurs. Important features of these data are 1) the longitudinal nature of the predictors, 2) the presence of strong correlations between antibodies that target the same protein and 3) high-dimensionality.

To tackle this prediction problem, we propose a modelling approach whereby the longitudinal trajectories of antibodies measuring the same protein are modelled using latent process mixed models, a multivariate extension of mixed models that allows us to deal with problems (1) and (2). Based on these models, we summarize the longitudinal information via the predicted protein- and antibody-specific random effects. Then, we employ the (high-dimensional) summaries thus derived as predictors in an elastic-net penalized Cox model, from which we derive predictions of time to LoA. We first compare the predictive performance of our model to that of simpler prediction strategies using simulated data, considering different scenarios that allow to understand how our approach can improve predictions. Then we apply our model to the MarkMD data, showing that predictions of time to LoA in DMD can be considerably improved by fully exploiting the available information on the dynamic evolution of a large number of biomarkers.

AUTHORS/INSTITUTIONS: M. Signorelli, R. Tsonaka, Biomedical Data Sciences, Leiden University Medical Center, Leiden, NETHERLANDS|P. Spitali, Human Genetics, Leiden University Medical Center, Leiden, NETHERLANDS
Abstract Body: Motivation: Calculation of the magnitude of treatment effects or differences between two groups is a common task in quantitative science. Effect sizes allow the quantification of influences of independent variables (features) on dependent variables (e.g., treatment outcomes). However, standard difference-based effect size measures fail to capture treatment-caused impacts on the data if the effects are not reflected by the central tendency. Moreover, the processing of big data in life sciences often implies a large number, e.g. thousands, of variables (features). In such a setting, visual inspection or manual analysis of all variables for their suitability to quantify treatment effects or group differences are unfeasible.

Results: We propose a novel non-parametric measure of effect size, “Impact”, obtained as the sum of two separate components, comprising (i) the change in the central tendency of the group-specific data normalized at the overall variability and (ii) the difference in the probability density of the group-specific data. Results obtained on artificial data and empirical biomedical data showed that by this additional component Impact outperformed Cohen’s d when used as a basis for feature selection. It is demonstrated that in a multivariate setting standard statistical analyses fail to identify effects that led to changes in the shape of the data distribution.

Conclusions: It is demonstrated that while classical effect size measures share the failure to observe an effect in changes in the shape of the data distribution, but not in the central tendency, with standard statistical analyses, the proposed effect size measure shares the ability to observe such an effect with machine learning algorithms. These have no problem in using the shape information for successful classification or group separation tasks. Therefore, the proposed effect size measure addresses a data science and machine learning context more than statistical data analysis.


Authors/Institutions: J. Lötsch, Goethe - Universität, Frankfurt am Main, GERMANY|J. Lötsch, IME-TMP, Fraunhofer Institute of Molecular Biology and Applied Ecology, Frankfurt am Main, GERMANY|A. Ultsch, DataBionics Research Group, University of Marburg, Marburg, GERMANY|
Abstract Body: An important problem in Statistics is the study of longitudinal data taking into account the effect of other explanatory variables such as treatments and time and, at the same time, incorporate into the model the time dependence between observations on the same individual. The latter is especially relevant in the case of having nonstationary correlations, as well as nonconstant variances for the different time points at which measurements are taken. Structured Antedependence (SAD) models constitute a well known commonly used set of models that can accommodate this behavior. These covariance models can include too many parameters and estimation can be a complicated optimization problem requiring the use of complex algorithms and programming. In this paper, a new Bayesian approach for analyzing longitudinal data within the context of structured antedependence models is proposed. This innovative approach takes into account the possibility of having nonstationary correlations and variances, and proposes a robust and computationally efficient estimation method for this type of data. We consider the joint modelling of the mean and covariance structures for the general structured antedependence (SAD) model, estimating their parameters in a longitudinal data context. Our Bayesian approach is based on a generalization of the Gibbs sampling and Metropolis-Hastings by blocks algorithm, properly adapted to the SAD models longitudinal data settings. Finally, we illustrate the proposed methodology by analyzing several examples where SAD models have been shown to be useful: the small mice, the speech recognition and the race data sets.

AUTHORS/INSTITUTIONS: V. Núñez-Anton, Department of Econometrics and Statistics (A.E. III), University of the Basque Country UPV/EHU, Bilbao, Bizkaia, SPAIN|E. Cepeda-Cuervo, Department of Statistics, National University of Colombia, Bogotá, COLOMBIA|
The causal effect of a treatment is of fundamental interest in the social, biological, and health sciences. Instrumental variable (IV) methods are commonly used to determine causal treatment effects in the presence of unmeasured confounding. In this work, we study a new IV framework with randomly censored outcomes where the causal treatment effect is quantified as complier quantile causal effect (CQCE). The CQCE is identifiable under weaker conditions than the complier average causal effect when outcomes are subject to censoring, and it can provide useful insight into the dynamics of the causal treatment effect. Employing the special characteristic of IV and adapting the principle of conditional score, we uncover a simple weighting scheme that can be incorporated into the standard censored quantile regression procedure to estimate CQCE. We develop robust nonparametric estimation of the derived weights in the first stage, which permits stable implementation of the second stage estimation based on existing software. We establish rigorous asymptotic properties for the proposed estimator, and confirm its validity and satisfactory finite-sample performance via extensive simulations. The proposed method is applied to a dataset from the Center for International Blood and Marrow Transplant Research to evaluate the causal effect of rituximab in diffuse large B-cell lymphoma patients.

AUTHORS/INSTITUTIONS: L. Peng, B. Wei, Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, UNITED STATES|M. Zhang, Medical College of Wisconsin, Milwaukee, Wisconsin, UNITED STATES|J. Fine, University of North Carolina-Chapel Hill, Chapel Hill, North Carolina, UNITED STATES|
A Method to use the dose-response relationship between exposure and disease for estimating the number of unexposed cases of a specific disease in a population

Abstract Body: Mortality statistics by country commonly provide disease-specific absolute and relative frequencies and rates of death by sex and age, but not by exposure status. However, it is often of interest to know how many of the diseased individuals, i.e., the cases, were exposed or not exposed to a specific risk factor.

For many risk factors a proportion of the population is unexposed, and among the exposed the distribution may be continuous, for example smoking or alcohol consumption. These variables are called semi-continuous or spike at zero variables. Information on the population distribution of those factors is often available from health surveys, however the proportion and absolute number of non-exposed deaths by cause is not known.

We present a method to estimate the proportion $p_{01}$ and subsequently the absolute number of exposed and non-exposed cases. The method requires an estimate of the exposure prevalence in the non-diseased population and an estimate of the relative effect of the exposure, i.e., a relative risk function if the exposure $X$ has a continuous distribution given $X>0$, or a categorical risk estimate for the exposure categories if the exposure is categorical. In the first instance, an estimate for $p_{01}$ is $p_{01} = p_{00} / (p_{00} + \int R(x) f(x) \, dx)$, where $p_{00}$ is the proportion of the non-exposed in the non-diseased, $R(x)$ is the risk function denoting the relative risk given dose $x$ compared to dose zero, and $f(x)$ the density function of $X$ in the non-diseased.

We provide theoretical justification for the method and present some theoretical examples for a given population distribution of $X$, such as log-normal or exponential distribution. We further suggest approaches for sensitivity analyses.

The method is applied to the estimation of the proportion and number of never smokers among lung cancer deaths in Germany. The sensitivity of the estimates to the underlying assumptions is discussed.

Authors/Institutions: H. Becher, Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, GERMANY|H. Becher, A. Aigner, Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, Berlin, GERMANY|
With the increasing availability of data in the public domains, there has been a growing interest in exploiting information from external sources to improve the analysis of smaller-scale studies. An emerging challenge in the era of big data is that the subject-level data are high-dimensional, but the external information is at an aggregate level and of a lower dimension. Moreover, heterogeneity and uncertainty in the auxiliary information are often not accounted for. In this paper, we propose a unified framework to summarize various forms of aggregated information via estimating equations and develop a penalized empirical likelihood approach to incorporate such information in logistic regression. When the homogeneity assumption is violated, we extend the method to account for population heterogeneity among different sources of information. When the uncertainty in the external information is not negligible, we propose a variance estimator adjusting for the uncertainty. The proposed estimators are asymptotically more efficient than the conventional penalized maximum likelihood estimator and enjoy the oracle property even with a diverging number of predictors. Simulation studies show that the proposed approach yields higher accuracy in variable selection compared with its competitors. We illustrate the proposed methodologies with a pediatric kidney transplant study.

AUTHORS/INSTITUTIONS: Y. Sheng, C. Huang, M. Kim, Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, UNITED STATES|Y. Sun, Epidemiology and Biostatistics, Columbia University, New York, New York, UNITED STATES|
Abstract Body: Context: Phase I dose-finding trials in oncology seek to find the Maximum Tolerated Dose of a drug under a specific schedule. Evaluating drug-schedules aims at improving treatment safety while maintaining efficacy. However, while we can reasonably assume that toxicity increases with dose, the relationship between toxicity and multiple schedules remains elusive.

Objective: The aim of this work was to develop a Bayesian dose-finding design for multiple schedules using PK/PD information to estimate the Maximal Tolerated Dose-Sequence (MTDS) at the end of the trial. We propose to model the binary toxicity via a summary of the profile of a toxicity-related biomarker, viewed as a continuous PD endpoint.

Methods: Firstly, the relationship between the PD biomarker profile and the dose-sequence is modelled using non-linear mixed models.

Secondly, we propose 2 Bayesian approaches to model the relationship between a summary of the biomarker profile and toxicity. For the first approach, we consider a Bayesian 2-parameters logistic model using a value of interest of the biomarker profile. For the second approach, we propose a Bayesian hierarchical model using a latent variable considering longitudinal values of the biomarker profile.

Finally, we integrated both models via simulations to analyse the entire relationship between the dose-sequence and toxicity in order to recommend the MTDS.

Results: We evaluated the operating characteristics of our methods through simulation studies under various scenarios. The results showed that our methods perform better than usual model-based designs in terms of percentage of MTDS correct selection. Moreover, due to the additional PK/PD information, our methods estimate more precisely the entire dose-sequence-toxicity curve and can propose untested sequences for expansion studies. Our methods will be applied to an ongoing dose-escalation trial for patients with relapsed or refractory acute myeloid leukemia (NCT03594955).

Conclusion: Our proposed dose-finding design for multiple schedules provides a reliable way to identify the MTDS when toxicity can be related to a PD biomarker. However, as the methods are applied at the end of the trial (once all data has been collected), they can be sensitive to the dose-escalation design. We are therefore extending our design to a sequential dose-allocation approach.

Mapping animal species abundance is important in order to better understand biodiversity loss, species migration and assessing the risk of spillover to human populations from zoonotic diseases. Multiple measurement tools are often used in order to obtain improved estimates of animal abundance. However, current approaches for combining multiple indices of abundance do not exploit the inferential benefits that might accrue from the joint spatial modelling of the different measurements. In this study, we have developed a class of multivariate generalized linear geostatistical models for modelling the cross-correlation in space among multiple imperfect metrics of animal abundance. We illustrate the development and application of the novel methodology in the context of a case study on Rattus norvegicus, a reservoir for leptospirosis in vulnerable urban communities in Salvador, Brazil. More specifically, we consider three indices of R. norvegicus abundance: rat signs, live traps and tracking boards. Our objective is to use the three outcomes in order to draw predictive inferences on a spatially continuous latent process which relates to R. norvegicus abundance and which we refer to as "rattiness". We then describe how the new methods can be used 1) to explore the association between "rattiness" and environmental factors, 2) to test for residual spatial correlation, 3) to evaluate the relative importance of each of the three outcomes in estimating "rattiness" and 4) to identify "rattiness" hotspots. We conclude by discussing both limitations and wider applicability of the developed methodology.
Many researchers find understanding p-values and confidence interval (CI) difficult and this is essential for the evaluation of results of any study. A test of hypothesis does not indicate what the difference is or how large it is. Simple statements in a study report such as ‘p < 0.05’ or ‘p = NS’ do not describe the results of a study well. Complementing the hypothesis test with a CI will indicate the magnitude of results and help researchers to decide whether the difference is of interest clinically. This study provides useful information for the interpretation of these statistical concepts to avoid misleading reports. The uses of these two statistical concepts and the differences between them are discussed on the basis of a comprehensive and selective literature search on the methods in scientific studies. The p-values are used to determine whether a null hypothesis is to be accepted or rejected and also enable the recognition of statistically significant findings. The CI provides an estimate of the precision with which a statistic estimates a population value. For instance, in a clinical trial of a placebo versus a hypotensive agent, each group with 10 patients, the change in blood pressure for the placebo was 17 mmHg and that of the hypotensive drug was 30 mmHg. If the pooled standard deviation were 15.5 mmHg by the two sample t-test: t = 1.9, df = 18 and p = 0.06. This fails to reach the conventional 5% significance level and may be declared not statistically significant. However, the potential benefit from a reduction in blood pressure of 13 mmHg is substantial and so the result should not be ignored. In this case, it would be misleading to state that ‘There was no significant difference between the drug A and B, and it would be better to quote the extra gain achieved of 13 mmHg, together with a 95% CI of −2 to 28 mmHg. In this way we can truly judge if the trial results are indicative of no difference or that, in a larger trial, the clinically important benefit of 17 mmHg indicated may be proven right. This enables conclusions to be made about the statistical significance and clinical importance of the study findings. This study therefore concludes that presentation of both the p-value and CI is desirable since they provide supplementary information. But if only one is to be reported, CI must be given a preference over the p-value for the sake of clinical significance.

AUTHORS/INSTITUTIONS:  K.A. Osuolale, A.Z. Musa, Biostatistics/Monitoring and Evaluation Unit, Nigerian Institute of Medical Research, Lagos, Lagos, NIGERIA}
Abstract Body: Historical data from previous clinical trials, observational studies and health records may be utilized in analysis of clinical trial data to strengthen inference. Instances of cases where historical data are crucial are clinical trials where assigning patients to a control group is considered unethical (oncology), patient accrual rates are low (rare diseases) or required sample sizes for detection of the actual treatment effects are unaffordable/infeasible (global health/nutritional interventions). Use of any external data should of course be conditioned on the validity of the source. More specifically, external data should only be used if it can be considered as a representative sample of the target population. Since the current data is considered the most immediate proxy to the target population, historical studies are used on the basis of their similarity to the current study. Under the Bayesian framework incorporation of information obtained from any other source than the current data is facilitated through construction of an informative prior. The existing methodology for defining an informative prior based on historical data relies on measuring similarity to the current data at the study level that can result in discarding individual patient data (IPD). This paper is focused on a family of priors that utilize IPD to strengthen statistical inference. IPD-based priors can be obtained as a weighted likelihood of the historical data where each individual's weight is reciprocal to their distance to the target population.

AUTHORS/INSTITUTIONS: S. Golchi, Biostatistics, McGill University, Montreal, Quebec, CANADA
Abstract Body: In the last few decades, there have been numerous researches to identify genetic variants associated with a single trait. However, it has turned out that relatively many variants are associated with more than one trait. They are called pleiotropic variants. Identification of pleiotropic variants can play a critical role in understanding missing heritability because they have not been revealed by single phenotype association studies. However, most of existing statistical methods for pleiotropic variants are limited to only quantitative traits. In this work, we propose new statistical approach to identify pleiotropic variants which can be associated with quantitative traits only, qualitative traits only or both. The proposed approach is to unify multiple elastic-net regularization models where variants selected by individual elastic-net models are summarized as their selection probability. In our simulation studies, we demonstrated that the proposed approach can select more pleiotropic variants with a small or moderate effect size than single phenotype association. We also applied the proposed approach to cowpea genotype data with 18 different quantitative and qualitative traits. We could find potentially pleiotropic variants missed by a single association study.

AUTHORS/INSTITUTIONS: K. Kim, H. Sun, Statistics, Pusan National University, Korea, Busan, KOREA (THE REPUBLIC OF) S. Wang, Biostatistics, Mailman School of Public Health, New York, New York, UNITED STATES
MODELING OF SURVIVAL OF HIV PATIENTS BY STAGES OF IMMUNO SUPPRESSION AND OPPORTUNISTIC INFECTIONS

Abstract Body: Modeling of survival of patients with respect to the occurrence of certain diseases have become of major concern or interest to many statisticians, especially in the field of Biostatistics for several years now. An important aspect of survival analysis is the estimation, evaluation, comparison and prediction of survival probabilities for individuals or patients. HIV infection has spread over the last thirty years (30) and has a great impact on health, welfare, employment and criminal justice sectors. People with advanced stage of HIV infection are vulnerable to secondary infections that are generally termed as Opportunistic Infections (OIs). This is because, these infections take advantage of the opportunity offered by a weakened immune system. The aim of this research is to investigate and model the survival, by stages of immune suppression and opportunistic infections on patients undergoing Antiretroviral Therapy (ART) in a population in South-South Nigeria. Four (4) different parametric models: Extreme, Lognormal, Logistics and Log-logistics distribution, and Cox semi-parametric model were considered in order to carryout comparison, model of survival, and prognostic factors influencing survival of patients. 221 HIV patients data obtained from St. Luke’s Hospital, Anua, for the period of 2008 to 2017 were used. An analysis was performed on this data in which comparison of aforementioned parametric models was done in order to obtain the best model using AIC and BIC as the goodness-of-fit criteria. The Extreme distribution had the lowest AIC and BIC value, indicating that Extreme distribution is the best fitted parametric model for modeling survival of HIV patients in the hospital. The Cox semi-parametric model indicates that Gender, CD4_count, Age and Number of Opportunistic Infections (Number of OIs) are significant prognostic factors that influences survival processes of HIV patients under the study, and that female suffers more HIV/AIDS related malignancies than the male.

AUTHORS/INSTITUTIONS: E.U. Umeh, M.E. Udofia, Statistics, Nnamdi Azikiwe University, Awka, Anambra, NIGERIA|
Penalized estimating equations for longitudinal cluster-randomized trials with missing continuous outcomes and small number of clusters

Abstract Body: Longitudinal cluster-randomized designs are frequently used for comparative effectiveness research in clinical trials. The methodologies for the three-level hierarchical design are not well understood, especially when the study aims to examine whether and how treatments are differently associated with the outcomes over time. The standard generalized estimating equations have been known to lead to an inflated type I error rate and bias due to the small number of available clinics and not completely random missing data in the longitudinal outcomes. We evaluate the performance of the inverse probability weighted (IPW) estimating equations on the inference of the treatment-by-time interaction effect with and without augmentation for missingness in continuous outcomes and individual-level treatment allocation mechanism combined with two bias-corrected variance estimators. We also propose penalized IPW and augmented IPW estimating equations using the smoothly clipped absolute deviation (SCAD) penalty. The results of the intensive simulation show an improved inference with bias corrections and high performance for the variable selection using the proposed methods. The results can be useful for choosing the appropriate analysis methods for the longitudinal cluster-randomization design. The proposed approaches are illustrated using the longitudinal cluster-randomized clinical trial of adults with serious mental illnesses.

Authors/Institutions: C. Kang, D. Zhang, Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States| J. Schuster, J. Kogan, C. Nikolajski, UPMC Insurance Services Division and UPMC Center for High-Value Health Care, UPMC, Pittsburgh, Pennsylvania, United States| C.F. Reynolds, Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania, United States
Abstract Body: The elastic net estimates can be interpreted as Bayesian posterior estimates when the regression parameters have a prior that compromises between Gaussian and independent Laplace (i.e., double-exponential) priors. A significant challenge in the elastic net is that it assumes that data are normally distributed, which makes it not robust to model misspecification. In this article, we propose a Bayesian semiparametric approach for an elastic net model that is based on empirical likelihood. This approach relaxes the normal assumption on data, and hence we avoid problems with model misspecification. Under the Bayesian empirical likelihood approach, the resulting posterior distribution lacks a closed-form and has nonconvex support, which makes the implementation of traditional Markov chain Monte Carlo methods such as Gibbs sampling and Metropolis-Hastings very challenging. To solve the nonconvex optimization and nonconvergence problems, we implement the Hamiltonian Monte Carlo approach.

AUTHORS/INSTITUTIONS: C. Moon, Statistical Science, Southern Methodist University, Dallas, Texas, UNITED STATES| A. Bedouin, Boehringer Ingelheim, Ingelheim, GERMANY
Abstract Body: We instituted a prospective registry to capture patient-reported clinical outcomes in a large, multi-site radiation oncology practice. Our aim was to explore use of inverse probability weighted propensity scores within artificial neural networks (ANN) to explore the association between clinical characteristics and adverse events (AEs) assessed by the Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE™). Methods: PRO-CTCAE questionnaires were administered at baseline, end-of-treatment, 3, 6, 12 months, then annually. Eligible patients were treated with radiation therapy with curative intent and completed selected PRO-CTCAE items based on cancer type. A patient was considered to have patient-reported treatment-related symptomatic AEs (PAE) if he/she had a score of 3 or 4 for any PRO-CTCAE item that was worse than baseline. ANN analysis with 4-fold cross-validation was utilized to compare likelihood of PAE by treatment modality and other characteristics with local interpretable model-agnostic explanations (LIME) to approximate feature weights. Due to disparate baseline characteristics between treatment modalities, we further compared use of ANN versus cross-validated logistic regression analysis (Logit) to create inverse probability weights in propensity score analyses (PS) to assess the effect of photon vs. proton therapy on PAE. Models were compared using areas-under-the-curve (AUCs). Results: 1,930 patients were eligible within 8 disease cohorts (PAE rates range 6.7%-58.3%). The AUC for modeling photon vs. proton therapy significantly increased from 0.73 (0.71-0.76) with Logit to 0.78 (0.76-0.80) with ANN, with LIME factor-weights identifying differences across treatment modalities in baseline characteristics. Subsequently, in PS analyses, the proton effect on PAE was significant with ANN (OR=0.79, p=0.045) when better accounting for baseline differences, but non-significant with Logit (OR=0.95, p=0.60). Conclusion: ANN is a powerful tool in predicting patient-rated AEs. Models are improved when employing multi-level modelling (i.e., better adjustments for confounding via PS), yielding more accurate estimates of treatment effect (e.g., significant benefit of proton therapy in reducing PAEs).
CONTROL ID: 3348474
TITLE: Randomization Tests for Multi-armed Randomized Clinical Trials

ABSTRACT BODY:

Abstract Body: This presentation contributes to the multiple comparisons problem in multi-armed randomized clinical trials. The view is taken that the linkage of the statistical test to the randomization procedure should be recognized to fulfill immediate utilities (e.g., statistical validity, selection of a randomization procedure in a trial) as well as to expand scientific and philosophical outlook in light of valid interpretation of statistical tests. In randomization tests, the model is constructed by the re-randomization of treatment assignments according to the randomization procedure, together with a test hypothesis. The model assumptions are ensured by the execution of the randomization procedure. In addition, the reference frame for data that might have been observed and the scope of the inference are defined. The goal of multiple comparisons centers around the control of the familywise error rate. The solution given by randomization tests depends on a conditional reference set. Its preservation of the type I error rate ensures that the various techniques for controlling the familywise error rate can be effective. We provide an efficient algorithm based on Monte Carlo simulation that makes computation of the test feasible, a demonstration of the preservation of type I error rate under heterogeneity, and an illustrative case study using data from multi-armed randomized clinical trials.

AUTHORS/INSTITUTIONS: Y. Wang, Medstar Health Research Institute, Washington, District of Columbia, UNITED STATES|W.F. Rosenberger, Statistics, George Mason University, Fairfax, Virginia, UNITED STATES|
Abstract Body: Clustered categorical data in the form of proportions often arise in toxicology, epidemiology, and other similar fields. For example, a set of toxicological data refers to live fetuses in a litter affected by treatment, and the number of live fetuses for each of two dose groups: control (C) and medium dose (M). Then the effect of the treatment (medium dose) can be quantified by the some well-known measures of association between a given disease and the treatment. The relative risk (RR) is one of three major useful measures of association for summarizing the results from such epidemiological cohort studies. This usually measures the relative change in the disease risk due to the application of the treatment. Standard approaches for estimating RR available in common software packages may lead to biased inferences when applied to correlated binary data. In this paper, based on the concepts of design effect and effective sample size used in sample surveys, we develop some simple and efficient inference procedures for estimating RR, by direct extensions of recently recommended methods for independent binary data. We also develop three more methods based on a hybrid method. The hybrid method simply combines two separate confidence intervals for two single risk rates to form a hybrid confidence interval for their ratio. Applying the extended modified Poisson approach to correlated data, we propose another asymptotic normality-based method. Moreover, we extend three existing methods recommended for complex survey data using different weighting schemes by using hybrid logarithmic transformation. Extensive simulation studies are conducted to investigate the performance of these proposed methods. Finally, a real-World example of toxicological study that motivated this research is used to illustrate the proposed methods.
Clinical trials of immune checkpoint inhibitor (ICI) therapy against CTLA4 or PD1/PDL1 suggest that immunotherapy of cancers is promising and the effect could be fairly durable; however, for most cancer types patients’ objective response rates to these treatments are less than 30%. Therefore, there is a critical need to identify patients who are most likely to benefit. Looking for biomarkers that can predict patients’ responses to ICI therapy is an important topic. Tumor mutation burden (TMB) and CD8+ T-cell abundance have been claimed to correlate with clinical benefit from ICI therapy; however, other immune cells have also been revealed to influence overall survival of cancer patients. In this study, we set out to address two questions by taking advantage The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC): 1) Could the relative abundance of immune cell subsets be correlated to the survival of cancer patients? 2) Could TMB be correlated to the abundance of certain immune cell subsets? To answer these questions, we employed \( \varepsilon \)-support vector regression (\( \varepsilon \)-SVR) as the core algorithm, along with the use of a loss function subject to the L1-norm penalty, to build an in silico cell composition deconvolution method. To construct the reference gene expression signature matrix for regression, a subset of differentially expressed genes were chosen from 166 microarray-based gene expression profiles for 10 types of immune cells by using ANOVA and minimizing condition number. In addition, a deep-learning approach based on autoencoder was used to reduce the dimensionality of features. Our method outperforms CIBERSORT in various benchmarks. Next, our cell composition deconvolution method was applied to estimate the immune cell subsets in the bulk gene expression profiles of TCGA LIHC. Then, the inferred immune cell compositions were associated with the survival time of TCGA LIHC patients by using the log-rank test. Our results reveal that a higher abundance of memory CD8 T cells in tumor microenvironment might be associated with a better survival rate in LIHC patients. In addition, TMB positively correlates to the abundance of naïve CD8 T cells, whereas it negatively correlates to the abundance of various subsets of immune cells with cytolytic activities.

**AUTHORS/INSTITUTIONS:** Y. Chiu, Y. Huang, Institute of Biomedical Informatics, National Yang-Ming University, Taipei City, Taiwan, TAIWAN| Y. Huang, Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei City, TAIWAN|
Model-based statistical methods for the dose-response analysis of toxicological and pharmacological mixtures.

Abstract Body: In our daily life, we are regularly exposed to dozens or even hundreds of chemicals (pesticide residues, food additives, medicines). Each of them may have adverse effects on living organisms but their safety is usually assessed only individually. There is therefore a need for statistical methods allowing scientists to assess the risk posed by the joint exposure to those chemicals.

Here we propose methods and a computer program to model the joint effect of chemicals from data. We assume that data at hand result from experiments where the effect of various dose combinations has been measured, for example in terms of average mortality in an animal population. The models considered belong to the class of parametric multivariate dose-response functions.

We review some of the most common classes of such functions, we outline the main issues encountered for parameter inference and model selection. We show that the widely used testing approach based on comparing a likelihood ratio to the quantile of a khi-square distribution is not grounded for datasets of common size. We outline how this issue can be sorted out by mean of Monte Carlo simulations. We also argue that in absence of well validated toxicological based on first principles, statistical models should be mostly focused on prediction accuracy.

The code developed during this work is made available as part of the R package MDR.

AUTHORS/INSTITUTIONS: G. Guillot, A. Macacu, Biostatistics, International Prevention Research Institute, Lyon, FRANCE
Applying the Self-Controlled Case Series method at the population level: Assessment of vaccination campaigns impact on yellow fever outbreaks in Africa

Abstract Body: Introduction: The self-controlled case series (SCCS) method is a case-only epidemiological study design for which individuals are used as their own control. As all time-invariant confounding are implicitly controlled for, this method is an alternative to classical cohort or case-control study designs when the risk of residual confounding is high. The SCCS method has successfully been applied at the individual level, but never at the population level. Here, we illustrate the use of the SCCS method at the population level by assessing the association between the occurrence of yellow fever (YF) outbreaks and the implementation of preventive mass vaccination campaigns (PMVCs) at the province level in Africa between 2005 and 2018.

Methods: Localization and date of YF outbreaks were identified from international epidemiological records, and information on PMVCs from international coordinators of vaccination activities. Using data collected in case provinces only, the incidence rate ratio (IRR) of YF outbreak between PMVCs-exposed and non-exposed time periods was estimated using a Poisson regression. The sensitivity to a range of assumptions was explored, and the results of the SCCS method were compared to those obtained based on a retrospective cohort study design. We further derived the number of YF outbreaks that have been prevented by PMVCs.

Results: The study sample consisted in 481 provinces in the 34 African countries endemic or at risk for YF. Over the 14-year observation period, we documented the implementations of PMVCs in 125 provinces; and the occurrence of 96 outbreaks in 81 provinces. Irrespectively of coverage achieved, exposure to PMVCs was found to reduce the risk for a province to experience an outbreak by 82% (95% CI: 63%-91%). This estimate was robust across a range of sensitivity analyses. We further conservatively estimated that PMVCs prevented 16.7 provinces (95% CI: 12.2-19.1) from experiencing YF outbreaks. Limitations of our study arise from the sparseness in available data on time-varying risk factors and surveillance quality.

Conclusion: Our estimates provide new evidence supporting PMVCs to prevent the risk of YF outbreak. To our knowledge, our study represents the first application of the SCCS method at the population level for a public health intervention evaluation.

Candidate biomarkers discovered in the laboratory need to be rigorously validated before advancing to clinical application. However, it is often expensive and time-consuming to collect the high quality specimens needed for validation; moreover, such specimens are often limited in volume. The Early Detection Research Network has developed valuable specimen reference sets that can be used by multiple labs for biomarker validation. To optimize the chance of successful validation, it is critical to efficiently utilize the limited specimens in these reference sets on promising candidate biomarkers. Towards this end, we propose a novel two-stage validation strategy that partitions the samples in the reference set into two groups for sequential validation. The proposed strategy adopts the group sequential testing method to control for the type I error rate and rotates group membership to maximize the usage of available samples. We develop analytical formulas for performance parameters of this strategy in terms of the expected numbers of biomarkers that can be evaluated and the truly useful biomarkers that can be successfully validated, which can provide valuable guidance for future study design. The performance of our proposed strategy for validating biomarkers with respect to the points on the receiver operating characteristic curve are evaluated via extensive simulation studies and compared with the default strategy of validating each biomarker using all samples in the reference set. Different types of early stopping rules and boundary shapes in the group sequential testing method are considered. Compared with the default strategy, our proposed strategy makes more efficient use of the limited resources in the reference set by allowing more candidate biomarkers to be evaluated, giving a better chance of having truly useful biomarkers successfully validated.
Abstract Body: Standard spatial cluster detection methods used in public health surveillance assign each disease case to a single location (typically, the patient’s home address), aggregate locations to small areas, and monitor the number of cases in each area over time. However, such methods cannot detect clusters of disease resulting from visits to non-residential locations, such as a park or a university campus. Thus we develop two new spatial scan methods, the unconditional and conditional spatial logistic models, to search for spatial clusters of increased infection risk. We use mobility data from two sets of individuals, disease cases and healthy individuals, where each individual is represented by a sparse sample of geographical locations (e.g., from geo-tagged social media data). The methods account for the multiple, varying number of spatial locations observed per individual, either by non-parametric estimation of the odds of being a case or by matching case and control individuals with similar numbers of observed locations. Applying our methods to synthetic and real-world scenarios, we demonstrate robust performance on detecting spatial clusters of infection risk from mobility data, outperforming competing baselines.

Authors/Institutions: R. Assuncao, R.N. Souza, W. Meira, Computer Science, UFMG, Belo Horizonte, Minas Gerais, Brazil; D. Neill, Center for Urban Science and Progress, New York University, New York, New York, United States
After new drugs are brought to market, it is necessary to track what kinds of adverse events (AE) occur because rare AEs may not be detected during clinical trials. Some organizations have been collecting information on suspected drugs and AEs via a spontaneous reporting system to conduct post-market drug safety surveillance. They use the information to detect a signal representing potential causality between drugs and AEs. The drug and AE data are often hierarchically structured, in which case the tree-based scan statistic can be used as a statistical data mining method for signal detection. Most of the AE databases have a large number of zero-count cells. In zero-count cells, not only an observational zero from the Poisson distribution but also a true zero exists. True zeros represent theoretically impossible observations or possible but unreported observations. The existing tree-based scan statistic assumes that all zero values as zeros observed from the Poisson distribution. Therefore, true zeros are not taken into account in the modeling, which can lead to bias in the inferences. In this study, we propose a tree-based scan statistic for zero-inflated count data in a hierarchical structure. Our simulation study has shown that in the presence of excess zeros our proposed tree-based scan statistic provides better performance than the existing tree-based scan statistic. We illustrate the two methods using the Korea Adverse Event Reporting System data of the Korea Institute of Drug Safety and Risk Management.

AUTHORS/INSTITUTIONS: G. PARK, I. JUNG, Yonsei University College of Medicine, Seoul, KOREA (THE REPUBLIC OF)
The DCV2 clinical trial (García et al., 2017) analyses the efficacy of a specific therapeutic vaccine as an alternative to current HIV treatment. A major challenge to analyze the time to viral rebound in the context of this trial consists of identifying convenient biomarkers. These have to be chosen among more than five thousand messenger RNAs (mRNAs). In addition, times to viral rebound in this trial are interval-censored.

We use an Accelerated Failure Time Model (AFTM) for the interval-censored times to viral rebound data including the mRNAs as predictors and consider an elastic-net penalization approach for the parameter estimation. The elastic-net method combines the penalizations from ridge regression and LASSO and allows automatic variable selection and continuous shrinkage, as well as the selection of groups of correlated variables. We derive the expression of the penalized likelihood function and attempt its maximization by means of an Expectation-Maximization (EM)-based algorithm.

The motivation for the use of the elastic-net approach to analyze the HIV data set is the fact that this method is not limited by a large number of predictors (> 5000 mRNAs) and the small number of observations (n=35). Furthermore, the elastic net method is appropriate when pairwise correlations are very high as it is the case with the mRNAs.

We apply this methodology to the data of the DCV2 trial. Preliminary analysis that used midpoint imputation for the interval-censored times to viral rebound and the elastic net penalization identified 4 mRNA from more than five thousand features. Our findings suggest that these 4 mRNAs (SMC4, B3GAT1, BTG1, MYCN) correspond to potential biomarkers for HIV viral rebound and may be considered to determine the effectiveness of the dendritic cell-based vaccine.

Furthermore, we plan to conduct a simulation study for different patterns of interval-censored times aiming to study the convergence of the EM-algorithm in these settings as well as the performance of the method using different numbers of correlated predictors.

Abstract Body: Nowadays, a huge amount of data is being collected all the time, so high-dimension databases are becoming very common to encounter. More specifically, with the advance of technology many biological information are now available at low costs -- data from genome, miRNA, mRNA, gene expression, protein, methylation, lipids, metabolism, phenotypes and so on. Many different studies have been done separately with each type of data, but more recently there is an increasingly interest in combine different data to gather more information. However, many classical methodologies used to this end assume the data matrix to be completed and numerical. Therefore, the heterogeneity of dataset with different variable types is not considered.

Alternatively, the Generalized Low Rank Models (GLRM) is a tool capable of dealing with large datasets of heterogeneous data. It is used for a single database, but can handle abstract data by using different loss functions, adequate to each variable type. GLRM is a very powerful tool that can deal with many different problems, but it is very recent, so its potential to work with multi-omics is yet unknown. In spite of this, some initial simulation studies showed that GLRM can deal with omics data well.

Thus, in this work, considering multiomics applications, we explore the possibilities of the GLRM for dimensionality reduction and prediction. Also, the latent structure of the problem is compared with other consolidated techniques. To allow a latent structure selection between different techniques, a model stability metric is implemented. Furthermore, an expansion for studies based on family data is proposed.

Keywords: multi-omics, generalized low rank models, matrix factorization, multivariate analysis, latent structure.

Acknowledgment: To CAPES (Brazil, Finance code 001) and FAPESP (Brazil, Process 2017/051257) for partial financial support.

AUTHORS/INSTITUTIONS: A.P. Vasconcelos, J.M. Pavan Soler, Statistics, Universidade de São Paulo, São Paulo, SP, BRAZIL|
Abstract Body: The sustainable development of bioeconomy involves the efficient use of natural resources including biomass and associated competitiveness for land area. The questions on how to predict the biomass availability in the future given the current tendency of land cover and future demand for biomass creates key challenges for the assessment of sustainability. Markov chain modeling, as a stochastic time-dependent process, can be one of the potential approaches to answer those questions. This paper presents the application of Markov chain modeling for numeric description of transition of land area under biomass production and prediction of land cover in the future. For the exemplary agricultural region of north-eastern Poland the historical and current data on land cover were collected according to four categories of land use (arable land, grassland, forests, and other land area). The time series for years 1990, 2000, 2006 and 2012 were completed on the basis of the linear raster 100 m resolution maps with the accuracy of data estimation 95-99%. The transition diagram of land use change assumes the four categories. This research provides results on practicality of the approach in prediction of the land cover transfers and potential biomass volumes in the future. It was concluded that instead of its limitation the Markov chain modeling can be recommended as a valuable supplementary tool in estimation of sustainability indicators related to land use and envisaged biomass production.
Abstract Body: Multimodal analysis combines and analyzes various types of brain images, and it has been widely applied in recent studies on brain diseases (Emine et al. 2019). In such analyses, dimension reduction is effective. One of the methods is the Supervised Multiblock Sparse Component Analysis (Kawaguchi et al. 2017). In this study, we propose the Sparse Nested Component Analysis developed from this conventional method.

Both methods consider two weights "within" and "between" images, and each brain image is compressed into each one-dimensional score by "within" weights, and the scores obtained from each brain image are combined as a total score by "between" weights. Therefore, "between" weights represent the contribution of each brain image in the total score. As in normal principal component analysis, multiple components of "within" weights are calculated in order, from the first component in the conventional method. Thus, complex high-dimensional data such as brain images are converted into one-dimensional scores that can be easily processed. These scores can serve as helpful biomarkers for disease diagnosis. We developed the proposed method based on an algorithm that can calculate multiple components of "between" weights for each "within" weight. Thus, it is expected that the more useful components for disease diagnosis can be extracted using the proposed method.

The proposed method was applied to real data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), which consists of a collection of datasets for all subjects with both of the two modalities; structural Magnetic Resonance Imaging (sMRI) and Diffusion Tensor Imaging (DTI) data. We implemented this method using the R software and calculated the scores. The proposed method provides useful scores for disease diagnosis that could not be obtained using the conventional method. In addition, it can be used for classification with reasonable prediction accuracy based on the receiver operating characteristic (ROC) analysis.

Reference


AUTHORS/INSTITUTIONS: R. Tajiri, A. Kawaguchi, Saga university, Saga, Saga, JAPAN|
Abstract Body: In species richness studies, citizen-science surveys where participants make individual decisions regarding sampling strategies provide a cost-effective approach to collect a large amount of data. However, it is unclear to what extent the bias inherent to opportunistically collected samples may invalidate our inferences. Here, we compare spatial predictions of forest ground-floor bryophyte species richness in Limburg (Belgium), based on crowd- and expert-sourced data, where the latter are collected by adhering to a rigorous geographical randomisation and data collection protocol. We develop a log-Gaussian Cox process model to analyse the opportunistic sampling process of the crowd-sourced data and assess its sampling bias. We then fit two geostatistical Poisson models to both data-sets and compare the parameter estimates and species richness predictions. We find that the citizens had a higher propensity for locations that were close to their homes and environmentally more valuable. The estimated effects of ecological predictors and spatial species richness predictions differ strongly between the two geostatistical models. Unknown inconsistencies in the sampling process, such as unreported observer’s effort, and the lack of a hypothesis-driven study protocol can lead to the occurrence of multiple sources of sampling bias, making it difficult, if not impossible, to provide reliable inferences.

AUTHORS/INSTITUTIONS: T. Neyens, C. Faes, CenStat, Hasselt University, Diepenbeek, BELGIUM|T. Neyens, L-BioStat, Catholic University of Leuven, Leuven, BELGIUM|N. Beenaerts, T. Artois, Centre for Environmental Sciences, Hasselt University, Hasselt, BELGIUM|P. J. DIGGLE, E. Giorgi, Lancaster Medical School, Lancaster University, Lancaster, UNITED KINGDOM|
Abstract Body: Separation or monotone likelihood can be observed in the fitting process of Poisson regression using maximum likelihood estimation (MLE) technique when one or more parameters diverge to infinity. The separation is very common in count data when sample size is small or there is huge number of zero count or there is sufficient number of strong predictors or mixture of two or more such conditions. The study investigates the consequence of separation in the standard Poisson models and provides a solution by incorporating Firth's (1993) type penalty term, which was originally proposed for bias reduction in MLE, in likelihood score equation that we may call penalized Poisson. The modified score equation guaranteed an achievement of convergence and finite estimate of the regression coefficient. An extensive simulation study was conducted to assess the performance of penalized Poisson model over standard Poisson and Zero-inflated Poisson in the presence of separation. Several simulation scenarios were considered for creating complete or quasi-complete or near-to-quasi-complete separation by varying the sample size, proportion of event in the binary predictor which make separation, and the magnitude of regression coefficient, log odds ratio, relating the binary predictor to the response. The results revealed that the penalized Poisson having profile penalized likelihood based confidence interval, performed better than the standard maximum likelihood based model, standard Poisson and Zero-inflated Poisson, in terms of bias, MSE and length of 95% confidence interval (precision) in all simulation scenarios. An illustration using real data also supported the simulation findings.
A Bayesian framework for case-cohort Cox regression

Abstract Body: The case-cohort study design (Prentice, 1986) is an increasingly popular approach for collecting cohort data. The infeasibility of obtaining certain covariates on the full cohort is circumvented by restricting the measurements to a randomly sampled subcohort, along with all remaining incident cases. This may be necessitated by time and cost constraints, as well as concerns over the wastage of valuable biological material. Existing proposals for analyzing case-cohort data are largely based on the Cox proportional hazards model. Weighted Cox regression approaches are the current norm in practice, motivated by the intuition that the oversampling of cases can be balanced by an appropriate overweighting of the subcohort controls. However, the data are used inefficiently, and inverse probability weights can be unstable. Approaches based on multiple imputation and nonparametric maximum likelihood have been proposed to address this, but suffer from incompatibility and computational issues respectively. We introduce a novel Bayesian framework for case-cohort Cox regression which avoids the aforementioned problems. Posterior sampling is carried out in two stages, where samples from the first stage serve as inputs in a pseudo-marginal MCMC algorithm (Andrieu & Roberts, 2009). The model for the baseline cumulative hazard function is nonparametrically specified and integrated out, allowing for only the log hazard ratio to be sampled. We illustrate the methodology with simulations, and apply it to the EPIC-Norfolk study to investigate risk factors for type-2 diabetes.

AUTHORS/INSTITUTIONS: A. Yiu, R. Goudie, B.D. Tom, MRC Biostatistics Unit, University of Cambridge, Cambridge, UNITED KINGDOM
Counterfactual mediation analysis with multistate model for surrogate and clinical time-to-event outcomes

Abstract Body: We introduce a counterfactual-based mediation analysis for surrogacy evaluation with time-to-event surrogate and clinical outcomes. Our approach accommodates censoring and competing risks. We use a multistate model for risk prediction to account for both transitions towards the clinical outcome and transitions through the surrogate outcome. We use the counterfactual framework to define the natural direct and indirect effects with a causal interpretation. Based on these measures, we define the proportion of the treatment effect on the clinical outcome mediated by the surrogate outcome. We estimate the proportion for both the cumulative risk and restricted mean time lost. We illustrate our approach using 18-year follow-up data from the SPCG-4 randomized trial of radical prostatectomy for prostate cancer. We assess time to metastasis as a surrogate outcome for prostate cancer-specific mortality.

Authors/Institutions: I. Weir, L. Trinquart, Biostatistics, Boston University School of Public Health, Boston, Massachusetts, UNITED STATES|J. Rider, Epidemiology, Boston University School of Public Health, Boston, Massachusetts, UNITED STATES|I. Weir, Center for Biostatistics in AIDS Research, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, UNITED STATES|
Evidence supporting the current World Health Organization recommendations of early antiretroviral therapy (ART) initiation for adolescents is inconclusive. We leverage a large observational data and compare, in terms of mortality and CD4 cell count, the dynamic treatment initiation rules for HIV-infected adolescents. Our approaches extend the marginal structural model for estimating outcome distributions under dynamic treatment regimes (DTR), developed in Robins et al. (2008), to allow the causal comparisons of both specific regimes and regimes along a continuum. Furthermore, we propose strategies to address three challenges posed by the complex data set: continuous-time measurement of the treatment initiation process; sparse measurement of longitudinal outcomes of interest, leading to incomplete data; and censoring due to dropout and death. We derive a weighting strategy for continuous time treatment initiation; use imputation to deal with missingness caused by sparse measurements and dropout; and define a composite outcome that incorporates both death and CD4 count as a basis for comparing treatment regimes. Our analysis suggests that immediate ART initiation leads to lower mortality and higher median values of the composite outcome, relative to other initiation rules.
In this talk, we introduce the entire procedure for response profiles analysis in observational longitudinal studies where time of intervention initiation varies with individuals. Generalized nonparametric piecewise regression models are developed to address changes in the mean response over time while accommodating the individual-specific time of intervention. Potential confounders on the relationship between time of intervention and response profiles can occur in observational longitudinal studies. The double-weighted estimation approach is proposed to access the mean response profile for individuals intervened at the same time and to explore the change of the mean response profile across times of intervention. The resultant estimator of the varying coefficients in the generalized nonparametric regression models is asymptotically consistent and normally distributed under the regularity conditions. The entire procedure is applied to provide the mean response profile of an inflammation biomarker of HIV-infected individuals in a guideline-based intervention study where antiretroviral therapy is initiated according to the World Health Organization guideline.
The differences among survival curves require testing and comparison. One measure for this is the restricted mean survival time (RMST), defined as the area under the survival curve up to a specific time point. It has several advantages compared with conventional analytic methods such as the log-rank test and Cox model. RMST can be estimated using the direct integration method of the estimated Kaplan–Meier curve (KM), the pseudo–observation method (PO), the inverse probability of censoring weighted method (IPCW), or flexible parametric modeling (FPM). In the past, Royston and Parmar (2011, 2013) recommended using FPM in the estimation of RMST, but its advantages over the other methods have not been adequately examined.

The operating characteristics of the four methods were investigated through extensive simulations under either proportional hazards (PH) or non-PH scenarios. The results showed, in brief, that the KM, PO, and IPCW methods provided unbiased estimators in all scenarios, while FPM indicated a small bias in some non-PH scenarios with delayed onset of treatment effect and/or a cured subset. FPM was also unstable in estimation for scenarios with very small numbers of events. FPM was a little better or equal in statistical power compared with the other methods in most scenarios. IPCW was lowest in power due to a larger standard error in scenarios with a small number at risk. The detailed results will be presented at the conference.

AUTHORS/INSTITUTIONS: R. Kimura, The University of Tokyo, Tokyo, JAPAN|K. Uemura, The University of Tokyo, Interfaculty Initiative in Information Studies, Tokyo, JAPAN|S. Nomura, National Cancer Center, Chiba, JAPAN|T. Hasegawa, SHIONOGI & CO., LTD., Osaka, JAPAN
Abstract Body: Prediction models are a useful tool to aid clinical decision making. For rare outcomes, there may be too few outcomes in a single dataset to reliably develop or validate prediction models. Therefore, combining individual participant data (IPD) from multiple studies for the purpose of developing and validating prediction models is particularly useful for conditions with rare outcomes. Benefits include having larger sample sizes with more outcomes than would have been possible with a single study, cost savings by reusing existing data, and the ability to evaluate a prediction model’s predictive performance across different centres, regions or countries. However, there are many challenges that come with combining IPD from different sources. These include assessing the quality of the data, differences in predictor and outcome definitions across studies, heterogeneity in the populations included across studies, and handling missing data including when predictors of interest were not recorded in all of the studies.

I will illustrate key challenges faced when combining IPD from different sources, and identify potential solutions and areas for further work. A detailed example will involve IPD from 78 studies for the purpose of developing and validating a prediction model for a woman’s risk of pre-eclampsia during pregnancy. IPD were used to externally validate existing published prediction models and develop new ones. Although 78 datasets were collected and harmonised, only 12 were used to develop new prediction models (1% of the available data), and case-mix differed across studies. The biggest challenge was missing data, where important predictors were not recorded in all individual studies. Studies also included different case-mix which can be both problematic and beneficial for evaluating generalizability of the model.

A key finding is that, although novel statistical methods can be used to impute missing predictors in a study by borrowing information from other studies, this can still be problematic in terms of convergence and applicability. Core outcome sets are already being developed (COMET), however core predictor sets are also needed. Researchers setting up new studies should think about the future use and reuse of their IPD, and ensure a core set of (known) predictors are collected as a minimum.

AUTHORS/INSTITUTIONS: K. Snell, R. Whittle, L. Archer, M. Hattle, J. Ensor, R. Riley, Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Keele, UNITED KINGDOM|J. Allotey, S. Thangaratinam, Barts Research Centre for Women’s Health, Queen Mary University of London, London, UNITED KINGDOM|J. Allotey, S. Thangaratinam, Pragmatic Clinical Trials Unit, Queen Mary University of London, London, UNITED KINGDOM
Abstract Body: Multiple System Atrophy (MSA) is a rare neurodegenerative disorder. Almost 80% of patients are disabled within 5 years of disease onset. The key pathogenic event when developing MSA is an abnormal accumulation of harmful proteins. Molecular causes and consequences of this aggregation need to be elucidated, for example using multiple omics datasets. We have access to DNA-methylome, miRNome, transcriptome and proteome data, measured in cell lines that show harmful protein aggregation and in negative controls. Standard sequential analysis of these data show no overlap of the significant genes. A combined analysis can detect relevant features shared by all datasets, improving the understanding of MSA.

Our aim is to develop a data integration method to identify consistent molecular biomarkers that can classify cells with protein aggregation across all datasets. Apart from the high dimensionality (p>N), also platform-specific heterogeneity between the omics data need to be considered. Several algorithmic approaches to integrate multiple datasets have been proposed, for example multi-group PLS (mg-PLS) and MINT. They decompose the datasets into joint and residual parts. The joint components capture consistent effects of the molecular measurements on the outcome across all datasets. The optimal components are obtained by iteratively maximising the covariance between the molecular measurements and a dummy matrix based on the binary outcome.

The drawbacks of mg-PLS and MINT are a lack of platform-specific parts in the decomposition, absence of a proper model for the binary outcome, and a risk of overfitting when data are high dimensional. Therefore, we propose a novel Probabilistic multi-group OPLS (mg-POPLS) model for multiple datasets in terms of joint, platform-specific and residual parts. Systematic differences between the omics data are incorporated in the model by including specific parts. The outcome is modelled via these components by using a latent probit model. The components and coefficients are estimated with maximum likelihood using an EM algorithm.

An extensive simulation study will be conducted to investigate the performance of mg-POPLS compared to mg-PLS. We apply the mg-POPLS method to the omics data measured in the cell lines to detect the most relevant features for separating MSA cases from controls.

Authors/Institutions: S. El Bouhaddani, H. Uh, J.J. Houwing-Duistermaat, Biostatistics and Research Support, University Medical Center Utrecht (UMCU), Utrecht, NETHERLANDS|J.J. Houwing-Duistermaat, Department of Statistics, University of Leeds, Leeds, UNITED KINGDOM|
Abstract Body: The workplace is a major vector for the spread of infectious diseases as influenza or gastroenteritis. It is also a place of development of emerging diseases related to working conditions and environment such as burnout or musculoskeletal disorders. Sick leave data are easily accessible to companies and an abnormal amount of sick leave spells can be the consequence of a high incidence of this type of phenomenon. Thus, the identification of sick leave outbreaks could be a useful contribution to help decision makers detect indirectly any potential changes that might affects employees’ physical or mental health.

In this context, we propose an adaptation of the Farrington algorithm [1],[2] for monitoring sick-leave data at the company level. The Farrington algorithm, used successfully in routine infectious disease surveillance systems, is based on a Quasi-Poisson regression model to predict potential outbreak after adjusting for linear trend and seasonality and underweighting past alerts. Our adaptation extends this model to consider also the within- and between-company characteristics (i.e. socio-demographic structure, multiplicity of sites, employees' insurance type, management). We developed a generalized linear model with covariates and random effect that leads to a specific computation of the threshold limit for the outbreak detection. We evaluated the model through extensive simulations and propose an application based on sick leave data from 1785 companies with more than 50 employees, followed for at least 6 years between 2010 and 2017.

Preliminary simulations provide promising results since it suggests that the performance of our model is close to the performance of the first Farrington model: we obtain a mean Probability of Detection of 58% and a False Positive Rate of 1.24% for outbreaks simulated as in [2]. This algorithm could then be used to help companies to detect drifts in sick leave but could also be used to address any surveillance issue in a multi-site setting.


Abstract Body: Background: Female Genital Mutilation/cutting (FGM/C) is a public health and human right issue, which is strongly anchored in customs and traditions, without any established benefit. The practice has both short and long-term consequences ranging from haemorrhage to death. It is estimated that about 200 million women and girls globally, have undergone FGM/C, with more cuttings being performed in Africa, the Middle East and Asia, with a report by UNICEF showing that in Africa, about 3 million girls are at risk of being cut each year.

In 2017, FGM/C prevalence among girls aged 0-14 years stood at 14.0% and 25.3% in Senegal and Nigeria, respectively. There is a huge geographical variation and a strong spatial structure of the practice requiring further examinations into where, when and how change is taking place using model-based statistical approaches to model, map and describe the characteristics of the hotspots where the practice is still rife.

Methods: Robust Bayesian Hierarchical spatial and spatio-temporal models, which simultaneously accounted for unobserved effects of space and time, as well as space-time interactions, whilst controlling for other linear and non-linear covariates were employed. These models were developed and fitted on the available datasets in a coherent mixed model regression framework. Posterior inference was carried out using Markov Chain Monte Carlo (MCMC) techniques, while model fit and complexity assessments utilised the Deviance Information Criterion (DIC) approach.

Results: There was an overall decline in the practice as found across the three countries. The Bayesian hierarchical modelling approach allowed us to jointly account for both individual-, household-, community-level factors, map and identify patterns and spatial variations in the practice, thus unmasking the hotspots across the three countries. Factors found to associate with higher risk of the practice included mother’s FGM/C status, support for FGM/C continuation, household wealth index, level of education of mother, region and type of place of residence, marital status and religion.

Conclusion: There has been a decline in FGM/C prevalence across the three countries, however, the menace of the practice is not yet over. More rigorous programmatic bespoke intervention approaches should be adopted and targeted on the identified hotspots.

AUTHORS/INSTITUTIONS: N. Kandala, Mathematics, Physics and Electrical Engineering, Northumbria University, Newcastle Upon-tyne, UNITED KINGDOM|N. Kandala, School of Public Health, University of Witswatersrand, Johannesburg, SOUTH AFRICA|
Abstract Body: We present the so-called COMBI method for evaluating genome-wide association studies which we have developed in prior work. In contrast to traditional locus-by-locus analyses, COMBI is a multivariate procedure which takes dependencies between different genomic loci into account. This is done by combining methods from machine learning and multiple testing. In a first stage of data analysis, a support vector machine (which is an inherently multivariate classification method) is trained. In a second stage, only the genomic positions with the largest contributions to the resulting classification rule are explicitly tested for association with the phenotype of interest, yielding a drastic dimension reduction. The thresholding of the association p-values for the selected positions is performed by means of a resampling procedure. Some remarks on software implementations of COMBI are made, and real data analyses are presented. The presentation is based on [1] and [2].

References:


ABSTRACT BODY:

Abstract Body: In studies with time-to-event outcomes, covariates of interest may also change over time. The classical Cox regression model can handle time-dependent covariates and assumes a constant effect on the log hazard function, which can be a limiting assumption in practice. In addition, when multiple time-dependent covariates are under study, it is also of great interest to model their joint effects by allowing a flexible functional form and to delineate their relative contributions to survival risk. Motivated by a cohort study investigating the effects of metabolic syndrome (MetSyn) on the risk of developing lung disease after a particulate exposure, we propose a partial single index hazards regression model with time-dependent covariates. We consider five components of MetSyn [body mass index, triglycerides, high density lipoproteins, glucose, and blood pressure]. The proposed method not only reduces the dimensionality of the covariates but also provides efficient estimates of the covariate effects. The flexible link function also allows nonlinear effects and interactions from the time-dependent variables on the log hazard function. We develop a two-stage iterative algorithm, in which a B-spline smoothing method is used to model the structured nonparametric single index component for the nonlinear covariate effects, followed by maximum partial likelihood estimation. We derive the asymptotic properties of the estimators for statistical inference. The proposed methods are illustrated using Monte Carlo simulation studies and applied to our cohort study.

Keywords: B-spline; nonparametric regression; partially linear single index model; time-dependent Cox regression model; metabolic syndrome; environmental exposures

AUTHORS/INSTITUTIONS: M. Lee, A.B. Troxel, M. Liu, Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, New York, UNITED STATES|S. Kwon, A. Nolan, Department of Medicine, Division of Pulmonary, Critical Care and Sleep Medicine, New York University, School of Medicine, New York, New York, UNITED STATES|A. Nolan, M. Liu, Department of Environmental Medicine, New York University, School of Medicine, New York, New York, UNITED STATES|
Abstract Body: Model selection in high-dimensional settings has received substantial attention in recent years, however similar advancements in the low-dimensional setting have been lacking. In this article, we introduce a new variable selection procedure for low to moderate scale regressions. This method repeatedly splits the data into two sets, one for estimation and one for validation, to obtain an empirically optimized threshold which is then used to screen for variables to include in the final model. In an extensive simulation study we show that the proposed variable selection technique enjoys superior performance compared to candidate methods, being amongst those with the lowest inclusion of noisy predictors while having the highest power to detect the correct model and being unaffected by correlations among the predictors. We illustrate the methods by applying them to a cohort of patients undergoing hepatectomy at our institution.

AUTHORS/INSTITUTIONS: M. Capanu, C. Begg, M. Gonen, Memorial Sloan Kettering Cancer Center, New York, New York, UNITED STATES|M. Giurcanu, University of Chicago, Chicago, Illinois, UNITED STATES|
Abstract Body: In 1972 Sir David Cox introduced to the Royal Statistical Society the proportional hazards model which has become perhaps the most widely used statistical model in medical research. Up until then, time-to-event data were handled either descriptively (Kaplan-Meier) or using specific parametric models for the distribution across time. By working with the hazard rather than the survival function a semiparametric model could be defined which handled covariates using a simple linear model but did not require a specified distribution across time. Software availability and close parallels to the generalized linear model made it increasingly popular through the 1990s for “censored data”.

Here we discuss several recent trends that challenge the unthinking use of proportional hazards. One of these is the desire to define a clinically interpretable estimand (ICH E9 draft addendum). Others question the causal interpretation of hazard ratio and the underlying plausibility of constant hazard ratio in a self-selecting population (those who have survived so far in this arm).

Finally we review the alternative approaches for use in clinical trials beyond the ICH E9 addendum, such as restricted mean survival time (RMST). We introduce the idea of a restricted hazard defined by dividing the cumulative distribution of event time by the RMST up to that specific time. This is clinically interpretable as it links two well known concepts to define a marginal hazard over a specific period, usually the length of the trial.

Authors/Institutions: J. Wei, M. Akacha, Novartis, Shanghai, CHINA|J.H. Roger, LSHTM, London, UNITED KINGDOM|
ABSTRACT BODY:

Abstract Body: In longitudinal data with many replications, a high-order autoregressive (AR) structure of the covariance matrix is required to capture the serial correlations between repeated outcomes. In this case, the high-order AR structure requires lots of parameters underlying the dynamic dependence of the data. Instead of the high-order AR structure for the covariance matrix, we propose an ARMA structured covariance matrix involving multivariate linear models. The covariance matrix is decomposed using autoregressive moving-average Cholesky decomposition (ARMACD) to account for the correlations between responses at each time point, within separate responses over time, and cross-correlation between different responses at different times. The ARMACD facilitates nonstationarity and heteroscedasticity of the covariance matrix, and the estimated covariance matrix is guaranteed to be positive definite. The proposed methods are illustrated using data derived from a study of nonalcoholic fatty liver disease.

AUTHORS/INSTITUTIONS: K. Lee, Department of Statistics, Sungkyunkwan University, Seoul, KOREA (THE REPUBLIC OF)
Abstract Body: Over the last few years, increasing attention has been given to first-hitting-time models, at least in the context of survival analysis. In biomedical applications, the idea is to model the health status as a stochastic process, for example a Brownian motion or a Gamma process, that degrades until it reaches a critical level (threshold), which may represent the death of a patient or the recurrence of a disease. The parameters of these processes (e.g., location and scale parameters in a Brownian motion) can depend on covariates, as well as the threshold. We develop a boosting algorithm to extend the use of first-hitting-time models to high-dimensional contexts. In particular, we focus on the situation in which low-dimensional clinical data must be combined with high-dimensional genetic data to build a prediction model. We show that the integration of these two sources of data in a first-hitting-time model is intuitive and avoids complicated weighting procedures. Finally, the novel approach is applied to a real data example.

Authors/Institutions: R. De Bin, V. Stikbakke, Department of Mathematics, University of Oslo, Oslo, Norway
An exploratory penalized regression to identify combined effects of functional agri-environmental variables

The development of new sensors allows observation at high frequency of the dynamics of agri-environmental variables affecting production. New challenge arises to model and predict fruit quality from these agri-environmental data. Modelling in agriculture had previously focused on plant growth and yield, whereas knowledge and models are not well established for fruit quality. There is a need for information extraction and statistical exploration in the case where the links between the quality variables and the functional environmental data are potentially non-linear. Usual functional data analysis (Ramsay and Silverman, 2005) are not adapted to extract information in the case of multivariate explanatory functional data. We propose a new approach using joint distributions of agri-environmental variables to explain a real (scalar) quality variable. Our exploratory approach is similar to boosting techniques that identifies various joint distributions associated to the explanatory environmental data. It associates penalized and structured regressions (Tibshirani and Taylor 2011, 2012) to select an optimal joint distribution that best explains the variable to be predicted. This approach has the additional advantage of being able to integrate, if necessary, so-called "expert knowledge" from the literature or other sources into the modelling process in order to improve the reliability of the results. Developed initially for agriculture, it is generic and can be used to solve scalar-on-function problems with the main hypothesis of identifying combined effects of functional explanatory variables. One limitation of this approach is the risk of overestimation, but various criteria are available to overcome it. The approach is validated on simulations and on real data.

G. GNANGUENON GUESSE, P. LOISEL, N. Hilgert, MISTEA, INRAE, Montpellier University, Montpellier, Occitanie, FRANCE|T. Simonneau, LEPSE, INRAE, Montpellier University, Montpellier, Occitanie, FRANCE|B. FONTEZ, MISTEA, Montpellier SupAgro, Montpellier University, Montpellier, FRANCE
Abstract Body: We propose a generalized linear model to predict a response trajectory in the future given individuals' characteristics at present in a longitudinal study. The coefficients in the generalized linear model vary as functions of both the present and the future time and hence illustrate the dynamic longitudinal profile of the future response. The bivariate time-varying coefficients are estimated by a novel nonparametric approach that takes advantage of features of both the kernel and the spline methods. The resulting coefficient estimator is asymptotically consistent under mild regularity conditions. We also develop a new bootstrap approach to construct simultaneous confidence bands for statistical inference about the coefficients and the predicted response trajectory based on the coverage rate of bootstrap estimates. We apply the proposed procedure to predict the probability trajectory of hypertension incidence given individuals' health condition in early adulthood and to examine the impact of risk factors in early adulthood on risk of the incidence of hypertension over age from the Framingham Heart Study.

AUTHORS/INSTITUTIONS: S. Kim, Statistics, Miami University, Oxford, Ohio, UNITED STATES|H. Cho, Biostatistics, University of Iowa, Iowa City, Iowa, UNITED STATES|M. Kim, Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, UNITED STATES
In clinical trials, there is considerable interest in investigating whether a treatment effect is similar in all patients, or that some prognostic variable indicates a differential response to treatment. To examine this, a continuous predictor is usually categorised into groups according to one or more cutpoints. Several weaknesses of categorisation are well known.

Objectives
To avoid the disadvantages of cutpoints and to retain full information, it is preferable to keep continuous variables continuous in the analysis. The aim is to derive a statistical procedure to handle this situation when individual patient data (IPD) are available from several studies.

Methods
For continuous variables, the multivariable fractional polynomial interaction (MFPI) method provides a treatment effect function (TEF), that is, a measure of the treatment effect on the continuous scale of the covariate (Royston and Sauerbrei, Stat Med 2004, 2509f). MFPI is applicable to most of the popular regression models, including Cox and logistic regression. A meta analysis approach for averaging risk functions across several studies has been proposed (Sauerbrei and Royston, Stat Med 2011, 3341f). A first example combining these two techniques (called Meta TEF) was published recently (Kasenda et al, BMJ Open 2016; 6:e011148). Another approach called meta-stepp was proposed recently (Wang et al, Stat Med 2016, 3704f). Using the data from Wang (8 RCTs in patients with breast cancer) we will illustrate various issues of our meta TEF approach.

Results and Conclusions
We used Meta TEF to investigate a potential treatment effect modifier in a meta analysis of IPD from eight RCTs. In contrast to cutpoint based analyses, the approach avoids several critical issues and gives more detailed insight into how the treatment effect is related to a continuous biomarker. Meta TEF retains the full information when performing IPD meta analyses of continuous effect modifiers in randomised trials. Early experience suggests it is a promising approach.
Abstract Body: Visualization algorithms have been widely used for intuitive interrogation of genomic data. Popularly used tools include MDS, t-SNE, and UMAP, among others. However, these algorithms are not tuned for the visualization of binary data and none of them take into account the hubness of observations for the visualization. In this presentation, I will discuss hubViz, our novel statistical model for hub-centric visualization of binary data. I will illustrate hubViz with simulation studies and its application to the single cell and bulk gene expression data, and the text mining data.

AUTHORS/INSTITUTIONS: D. Chung, Department of Biomedical Informatics, Ohio State University, Columbus, Ohio, UNITED STATES|J. Nam, School of Pharmacy, Sungkyunkwan University, Suwon, KOREA (THE REPUBLIC OF)|I. Jin, Department of Applied Statistics, Yonsei University, Seoul, KOREA (THE REPUBLIC OF)
Abstract Body: The analysis of categorical time series has been used to explore human memory patterns. Thus, for example, in an experiment of mentally tossing a coin, people believe, erroneously, that the coin alternates from heads to tails more often than really does. We present a study involving generation of random binary sequences by older adults either with Mild Cognitive Impairment (MCI) or with Alzheimer’s disease (AD). The results are compared with the control group (healthy older adults aged over 60). We conducted an experiment in which individuals were asked to mentally simulate a fair coin. To that end, the subjects were each to produce a single sequence of 60 head-tail outcomes, simulating the behaviour of a fair coin, without seeing (but perhaps remembering) the past outcomes. The study presented here is framed in the context of longitudinal data analysis in which a binary response, for a same individual, is repeated at 60 time points. A Markov chain of order (memory) k, taking values in a finite state space, is a well-known probabilistic model usually used to describe long memory processes (Baena-Mirabete and Puig, 2018). For more sophisticated processes, a combination of more than one Markov chains is a powerful alternative. In activities concerning the generation of random values by humans, one would expect to explain the deviation from randomness by a finite mixture that captures the population heterogeneity of the transition probabilities. We propose mixture models based on Markov chains from different approaches, as described in Baena-Mirabete and et al. (2019). We analyse whether there are significant differences in the binary sequences mentally generated by older adults with neurodegenerative disease compared to control group. The proposed models provide a tool that can be applied for early detection of Mild Cognitive Impairment and Alzheimer’s disease.

References


AUTHORS/INSTITUTIONS: S.B. Mirabete, Dept. de Matemàtiques, Universitat Autònoma de Barcelona, Barcelona, SPAIN|S.F. Guinea, Dept. de Psicologia Experimental, Universidad Complutense de Madrid, Madrid, SPAIN|R. Garcia-Viedma, Dept. de Psicologia, Universidad de Jaén, Jaén, SPAIN|E. Garcia, Dept. de Psicologia Experimental, Universidad Complutense de Madrid, Madrid, SPAIN|A. Junquera, Dept. de Psicología Experimental, Universidad Complutense de Madrid, Madrid, SPAIN|P. Puig, Mathematics, Universitat Autònoma de Barcelona, Cerdanyola del Vallés, Catalonia, SPAIN
Abstract Body: The Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI) provides a rich source of information on the incidence and survival rates of cancer in the United States. SEER has been used extensively to evaluate cancer health disparities. Individual cases captured in the registry data, however, have many missing information on important risk factors. The amount of missing information can vary between subgroups and change over time. The observed breast cancer incidence trends may be deceptive and inaccurate due to missing information for these factors.

In this project, we focus on examining breast cancer incidence trends by Estrogen Receptor (ER) status using data from SEER program during 1992-2017. The primary objective is to investigate whether using multiple imputation methods to impute missing values on breast cancer ER status corrects for bias in breast cancer incidence trends. When analyzing missing data, the assumption of missing at random (MAR) is generally used for imputation purposes. In contrast, the missing not at random (MNAR) hypothesis is rarely explored. In this study, we utilize multiple imputation methods to impute missing values for ER status with both MAR and MNAR assumptions. We then compare observed and model-imputed incidence trends for breast cancer by ER status to identify whether imputation changes incidence trends by a statistically significant degree. To rule out the possibility that changes in the trends may be due to non-imputation related changes, we also conduct a sensitivity analysis on all cases.

AUTHORS/INSTITUTIONS: P. Hu, Division of Cancer Prevention, U.S. National Cancer Institute/National Institute of Health, Bethesda, Maryland, UNITED STATES
Abstract Body: Recent technological advancements in neuroimaging make large amounts of brain imaging data available. However, extraction of meaningful information remains challenging due to the sheer volume and complexity of the data. Hence, there is a pressing need for statistical procedures that are computationally scalable and can accurately capture the neuronal structures from brain imaging data. We propose a fast algorithm for estimating the fiber orientation distribution (FOD) based on diffusion MRI data. This procedure treats the observed diffusion MRI signal at each voxel as a convolved and noisy version of the underlying FOD, and utilizes the spherical harmonics basis for representing FOD. The proposed method efficiently resolves the noise-amplification issue associated with this ill-posed inverse problem through appropriate regularization. Specifically, the coefficients at each level of spherical harmonics are shrunk by using a James-Stein type nonlinear shrinkage function adapted to this block of coefficients. This procedure significantly suppresses the noise and yields an estimator with low L2 risk. To further improve the estimation accuracy, a post estimation one-step super-resolution sharpening process is employed to enhance the localized peaks of the estimated FOD. The estimated FODs are used as input to a peak detection algorithm and the derived directions are then fed into a fiber tracking algorithm for reconstruction of the white matter fiber tracts. Subsequently, various brain structural related features (e.g., number of fibers) are derived. We illustrate the proposed method using both synthetic data and data from the Human Connectome Project (HCP). Moreover, we investigate the relationships between gender and handedness with brain structural connectivity features based on the HCP young adults data. We find significant gender main effects on the number of fibers within five subcortical regions. We also find significant gender-handedness interaction effects in two sub-cortical regions (Hippocampus and Amygdala).

Authors/Institutions: S. Hwang, Biostatistics, University of California, Davis, Davis, California, UNITED STATES|T. Lee, D. Paul, J. Peng, Statistics, University of California, Davis, Davis, California, UNITED STATES|
Abstract Body: Greater understanding of the pathways through which an environmental mixture operates is important to design effective interventions. We present new methodology to estimate natural direct and indirect effects and controlled direct effects of a complex mixture exposure on an outcome through a mediator variable. We implement Bayesian Kernel Machine Regression (BKMR) to allow for all possible interactions and nonlinear effects of (1) the co-exposures on the mediator, (2) the co-exposures and mediator on the outcome, and (3) selected covariates on the mediator and/or outcome. From the posterior predictive distributions of the mediator and outcome, we simulate counterfactuals to obtain posterior samples, estimates, and credible intervals of the mediation effects. Our simulation study demonstrates that when the exposure-mediator and exposure-mediator-outcome relationships are complex, BKMR–Causal Mediation Analysis performs better than current mediation methods. We applied our methodology to quantify the contribution of birth length as a mediator between in utero co-exposure to arsenic, manganese and lead, and children’s neurodevelopmental scores, in a prospective birth cohort in Bangladesh. Among younger children, we found a negative (adverse) association between the metal mixture and neurodevelopment. We also found evidence that birth length mediates the effect of exposure to the metal mixture on neurodevelopment for younger children. If birth length were fixed to its 75th percentile value, the harmful effect of the metal mixture on neurodevelopment is attenuated, suggesting nutritional interventions to help increase fetal growth, and thus birth length, could potentially block the harmful effect of the metal mixture on neurodevelopment.

Abstract Body: In brain network analysis, nodes are usually defined by atlas-based regions of interest (ROIs), and a correlation is a widely used similarity measure for functional connections. However, we sometimes need data-dependent nodes to investigate human brain activity specified in a given data.

In this study, we propose a simple and fast linear matrix factorization method that works well for small amount of high dimensional data, called kernel-based spatial map decomposition (KSMD). The KSMD is a kernel-based singular value decomposition with simultaneous diagonalization which extracts orthonormal spatial maps and their features from an auto/cross-correlation matrix. When a given uni/multi-modal brain imaging data has high dimension and small observations (small n, large p data), this kernel trick on the correlation matrix reduces from \( O(p^2) \) space and \( O(p^3) \) time complexity to \( O(n^2) \) and \( O(n^3) \), respectively (\( p \gg n \)).

After obtaining spatial maps, we detect active modes in the spatial map by finding persistent critical level sets during Morse filtration. Until now, highly activated brain regions which have larger values than a predefined threshold in the spatial map are mainly visualized and interpreted. The proposed method can approximate the spatial maps by finding active modes without determining the proper threshold.

In experiments, we applied the proposed methods to three datasets: (1) multi-modal brain network based on simultaneous FDG PET and fMRI obtained from healthy normal subjects from 20s to 70s, (2) multi-modal brain network based on FDG PET and MRI of attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD) children, and pediatric controls (PedCon), and (3) uni-modal brain network based on FDG PET in Alzheimer’s disease neuroimaging initiative (ADNI) database consisting of normal controls, mild cognitive impairments nonconverter (MCInc), MCI converter, and Alzheimer’s disease. The results showed that the features obtained by the proposed uni/multi-modal linear method had better performance in age prediction and group classification than the existing multi-modal linear method, group independent component analysis (GICA). Moreover, the proposed active mode detection method could make it easy to visualize and interpret the spatial maps corresponding to the features with good performance.

AUTHORS/INSTITUTIONS: H. Lee, H. Choi, E. Kim, B. Kim, D. Lee, Seoul National University Hospital, Seoul, KOREA (THE REPUBLIC OF)|H. Kang, Y. Huh, B. Kim, D. Lee, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|
CONTROL ID: 3361789

TITLE: Semiparametric Accelerated Failure Time Modeling for Multivariate Failure Times under Multivariate Outcome-Dependent Sampling Designs

ABSTRACT BODY:

Abstract Body: An outcome-dependent sampling (ODS) design, a retrospective sampling scheme where one observes the covariates with a probability depending on the outcome and selects supplemental samples from the most informative and appealing segments, is a cost-effective study design. In practice, multivariate data have increasingly been encountered in many contexts. It might be of interest to consider several disease outcomes or several subtypes of a disease simultaneously.

In this paper, we propose a statistical inference procedure for fitting multivariate failure time data from multivariate-ODS designs. We consider a semiparametric accelerated failure time (AFT) model, which directly links the failure time to covariates through a log function without specifying the error distribution. We take a marginal model approach to handle the correlated feature among failure times and incorporate the inverse of sampling probability weights to cover the biased sampling feature of the multivariate-ODS design. The proposed estimators are shown to be consistent and asymptotically normal. Extensive simulation studies suggest that our proposed design and estimator are more efficient than other competing estimators based on simple random samples. We apply the proposed methods to analyze the Busselton Health Study data.

AUTHORS/INSTITUTIONS: T. Lu, Mathematics, National Taiwan Normal University, Taipei, TAIWAN|S. Kang, Applied Statistics, Yonsei University, Seoul, KOREA (THE REPUBLIC OF)|H. Zhou, Biostatistics, Univ. of North Carolina at Chapel Hill, Chapel Hill, North Carolina, UNITED STATES|
ABSTRACT BODY:

Abstract Body: Under-diagnosis has been reported for many disease phenotypes with varying degree by health care institutions, physicians and patient characteristics. Electronic health records (EHRs) contain patient information on diagnosis, symptoms, laboratory tests, prescriptions, etc, thereby enabling researchers to assess under-diagnosis and identify responsible factors. Building on a phenotyping method that we recently developed by exploiting the concept of “anchor variables”, we propose a novel method for testing homogeneity of disease under-diagnosis across sub-sections in an EHR system. In our method, the degree of under-diagnosis is defined by the proportion of diagnosed patients among all patients having the disease of interest in an EHR. We propose a variance-component score test to assess whether such proportion varies across EHR sub-sections using a generalized linear mixed effect model framework. Our method necessarily addresses the challenge that the under-diagnosed patients and those without the phenotype are mingled together. To this end, assuming that all the diagnosed cases can be effectively identified, we validate a subset of the remaining patients in the EHR regarding their true phenotype status. We evaluate the performance of the test through extensive simulation studies and application to Penn Medicine EHR to assess under-diagnosis of several disease phenotypes.

AUTHORS/INSTITUTIONS: E. Bae, J. Chen, Department of Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES|Y. Ma, Department of Statistics, Penn State University, College Park, Pennsylvania, UNITED STATES|D. Rader, Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES|D. Herman, Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, UNITED STATES
Is the time to progression ratio an appropriate endpoint for clinical trials? A critical examination of the current practice and suggestions for a new methodology.

Abstract Body: The time to progression ratio (TTPr) is a novel endpoint in Phase I/II oncology trials, which is frequently applied to evaluate the efficacy of molecular targeted treatments in late stage cancer. The general idea of the design is that a patient serves as their own control.

To calculate the TTPr for an individual, the time to progression (TTP) under the experimental targeted treatment is divided by the last TTP under standard treatment. If the TTPr exceeds a certain value (typically 1.3), the person is considered a responder. Subsequently, a binomial test is performed, investigating if the proportion of responders is significantly higher than a certain threshold (typically 15%).

In this work, the current practice for the TTPr is critically examined. Using elementary calculations and simulations based on reasonable assumptions, we point out numerous shortcomings of the current methodology. Notably, the applied threshold values will often lead to rejection of the null hypothesis even if the experimental treatment is harmful. On the other hand, the approach features little power under appropriately chosen thresholds.

As a remedy to these shortcomings, we propose a new methodology for evaluating trials in which patients serve as each owns control. This methodology enables the formulation of meaningful null hypotheses and markedly outperforms the current approach in terms of power.

Authors/Institutions: D. Edelmann, T. Terzer, T. Hielscher, A. Benner, German Cancer Research Center, Heidelberg, GERMANY
Information for statistical analysis is frequently available in different micro databases. Each data base contains some of the variables of interest. Statistical matching aims at combining information obtained from different non-overlapping sample surveys referred to the same target population.

Formally, let $(X, Y, Z)$ be a random variable with joint discrete distribution $P$.

Furthermore, let $A$ and $B$ be two independent samples of $n_A$ and $n_B$ independent and identically distributed records from $(X, Y, Z)$. Assume that $(X, Y)$ are observed in sample $A$ while $(X, Z)$ are observed in sample $B$. The main goal of statistical matching, at a macro level, consists in estimating the joint distribution of $(X, Y, Z)$ from the samples $A$ and $B$. Due to the lack of joint information on $Z$ and $Y$ given $X$, such a distribution is not identifiable (identification problem). In other words, the sample information provided by $A$ and $B$ is actually unable to discriminate among a set of plausible models for $(X, Y, Z)$, leading to uncertainty about the data generating model.

In order to overcome the identification problem, alternative techniques have been proposed in the literature: (i) techniques based on the conditional independence assumption between $Y$ and $Z$ given $X$ (CIA); (ii) techniques use external auxiliary information regarding the statistical relationship between $Y$ and $Z$. However, it is possible that neither case is appropriate: the CIA assumption is rarely met in practice and external auxiliary information is hardly ever available.

In this paper we propose the use of Bayesian networks (BNs) to deal with the statistical matching scenario in the identification problem framework for categorical data.

The use of BNs is motivated by the following advantages: (i) BNs are widely used to describe dependencies among variables in multivariate distributions; (ii) BNs admit convenient recursive factorizations of their joint probability useful both for parameters estimation and for uncertainty evaluation in a multivariate context. In particular, the modularity of the graphical model allows to separately deal with: 1) subgraphs induced by nodes (variables) belonging to the same sample; 2) subgraphs induced by variables observed on different samples. In other words, those parameters affected by uncertainty are separated from those directly estimable from the available sample information. In this way also computational complexity is limited to subsets of variables.

**AUTHORS/INSTITUTIONS:** P. Vicard, Economics, University Roma Tre, Rome, ITALY|D. Marella, University Roma Tre, Rome, ITALY|V. Vitale, Sapienza Università di Roma, Rome, ITALY
The estimation of survival probability is of interest as a measure of absolute risk. For example in the context of ageing using data from the Leiden longevity study (400 nonagenarian sibships), genotype-specific survival probability can be estimated [1]. However, for clustered data, these estimates might be biased if the frailty distribution is not correctly specified. This study first aims to investigate the effect of frailty misspecification on the estimation of regression coefficient, baseline hazard parameters, and survival probabilities via Monte Carlo simulations. Clustered survival data is generated assuming a parametric baseline hazard with varying number of clusters, frailty variance, and frailty distribution. A wrong frailty distribution is then fitted to the data.

The simulation results suggest that misspecification of the frailty distribution may lead to biased estimates of regression coefficient, baseline hazard parameters, or survival probability. The number of clusters does not seem to have an effect on the estimation of parameters which means that the misspecification effect cannot be compensated by using more clusters. In most cases, the baseline hazard scale parameter corrects for the wrong frailty. However, for some extreme cases, it appears that the baseline hazard scale parameter cannot adjust and thus the regression coefficient is affected. This suggests that when using a flexible baseline hazard e.g. via splines might solve this problem and improve the estimation of regression coefficient and survival probability. This will be investigated both via Monte Carlo simulations and by application to Leiden longevity study data.

References

AUTHORS/INSTITUTIONS: A.M. Muli, A. Gusnanto, J.J. Houwing-Duistermaat, Statistics, University of Leeds, Leeds, UNITED KINGDOM| M. Rodríguez-Girondo, Leiden University Medical Center, Leiden, NETHERLANDS| J.J. Houwing-Duistermaat, Alan Turing Institute, London, UNITED KINGDOM| J.J. Houwing-Duistermaat, Utrecht University Medical Center, Utrecht, NETHERLANDS|
Multi-arm trials are an efficient way of simultaneously testing several experimental treatments against a shared control group. As well as reducing the sample size required compared to running each trial separately, they have important administrative and logistical advantages. There has been controversy over whether multi-arm trials should correct for the fact that multiple null hypotheses are tested within the same experiment. Previous opinions have ranged from no correction being required, to a stringent correction (controlling the probability of making at least one type I error) being needed. We propose that controlling the false-discovery rate (FDR) provides a suitable compromise, with an appealing interpretation in multi-arm clinical trials.

We investigate the properties of the different correction methods in terms of the positive and negative predictive value (respectively how confident we are that a recommended treatment is effective and that a non-recommended treatment is ineffective). We show that controlling the FDR provides good properties. It retains the high positive predictive value of FWER correction in situations where a low proportion of treatments is effective. It also has a good negative predictive value in situations where a high proportion of treatments is effective. In a multi-arm trial testing distinct treatment arms, we recommend that sponsors and trialists consider using the FDR.
Probabilistic models of mutational catalogues and statistical methods for the empirical evaluation of tools for the discovery of mutational signatures

Abstract Body: Mutational processes generate unique patterns of somatic mutations referred to as “mutational signatures”. Accordingly, the distribution of somatic mutations found in a cancer genome, the so-called “mutational catalogue”, is given by the superposition of the signatures of all the active processes, each with a specific intensity [1]. A great amount of effort has been put in formalizing these notions and in developing statistical methods for the discovery of signatures. Signatures and catalogues are usually defined as probability distributions over a domain of mutation types. Starting from a sample of mutational catalogues, de-novo signature discovery consists in expressing each catalogue as a linear combination of basic signatures with positive weights corresponding to the intensity of each process. New methods are being regularly published to solve this deconvolution problem in which both signatures and weights have to be estimated. Most of these methods are based on non-negative matrix factorization algorithms and their probabilistic versions and have been implemented in R. To date, the applications of these methods to tens of thousands of genomes have led to the identification of tens of signatures for most cancer types.

Despite the advances on these topics, little effort has been put on the empirical comparison of the existing algorithms and packages. In an article that have just been published [2], we have introduced a new probabilistic model based on the zero inflated Poisson distribution to simulate mutational catalogues from fixed signatures and used it to evaluate the principal algorithms in a comprehensive simulation study. In the present communication, we aim at introducing this model and a more recent improvement based on the zero inflated negative binomial distribution. Moreover, we propose appropriate metrics to evaluate the performance of methods for signature discovery. While the general idea consists in comparing the “true” signatures used for simulations with the estimated ones, the non-supervised learning nature of the problem makes this task particularly slippery. An R package that will make it possible to carry out such simulation studies is under development.


Authors/Institutions: V. Perduca, Laboratory of Applied Mathematics MAP5, Université de Paris, Paris, France
H. Omichessan, G. Severi, V. Perduca, Centre de recherche en Epidémiologie et Santé des Populations, INSERM, Villejuif, France
Classification is a challenging problem in genomics, where the number of features often outnumbers the number of observations. To overcome the challenges associated with high-dimensionality, classifiers generally employ some form of regularization (e.g. the independence rule, sparsity-inducing penalties, feature selection, dimension reduction). These forms of regularization essentially discard or disregard aspects of the data. For example, feature selection methods impose a working sparsity assumption and aim to select a small subset of relevant features while ignoring many null features. However, genomic data rarely satisfy these assumptions, due to the presence of various unobserved biological variables in addition to the variable of primary interest. These latent variables give rise to dense latent signals, and in their presence, the aforementioned strategies for coping with high-dimensionality may be suboptimal.

To avoid discarding potentially predictive information, we propose the cross-residualization classifier (CRC), an ensemble classifier which uses the latent signals in addition to the signal of primary interest. The CRC is shown to generally outperform popular classifiers on a wide variety of genomic datasets. However, it is computationally burdensome. Fitting the CRC involves multiple rounds of leave-one-out computations on the full feature matrix. Specifically, leave-one-out computations must be performed to (1) adjust the latent signals from the training observations and (2) fit the classifiers in the ensemble, one of which is a regularized linear discriminant classifier with a penalty parameter that must also be estimated in a leave-one-out fashion.

We derive and implement a computationally efficient algorithm for fitting the CRC, using nontrivial matrix downdating techniques. Our implementation is shown to be significantly faster than the naive implementation and has running time comparable to classifiers commonly used in genomic applications, as illustrated by simulations on datasets of varying sizes.

AUTHORS/INSTITUTIONS: Y. Pan, J.A. Gagnon-Bartsch, Statistics, University of Michigan, Ann Arbor, Michigan, UNITED STATES
Abstract Body: Continuous treatments (or named exposures) arise frequently from observational studies in which the study population could be potential with heterogeneity. Identifying and estimating heterogeneous causal effects from such observational data gains great attention when investigating health policy-relevant questions, such as understanding the vulnerable subpopulation in health policy research. Causal inference literature tackling this question has been restricted to binary exposure allocation, ignoring the fact that the exposure can be naturally continuous in many research fields. In this paper, we develop an empirical risk minimization (ERM) based approach named C-Learner to estimate the non-parametric heterogeneous exposure-response function (ERF) of a continuous exposure via flexible supervised learning methods in statistics and machine learning. We provide the identification conditions of the causal estimand and derive theoretical properties of the proposed estimator based on (adaptive) varying coefficient models with kernel smoothing, in which we show a non-parametric version of the quasi-oracle property. The algorithm is implemented based on a flexible two-step implementation framework in various simulation settings. A practical consideration is illustrated by analyzing the Medicare population across the Northeastern US to estimate the heterogeneous causal effects of PM$_{2.5}$ on mortality across subpopulations in different geographic regions. An available R package cle learner provides an easy-to-use implementation of our method.

AUTHORS/INSTITUTIONS: X. Wu, X. Li, F. Dominici, Biostatistics, Harvard University, Boston, Massachusetts, UNITED STATES|A. D'Amour, Google Research, Cambridge, Massachusetts, UNITED STATES|
Abstract Body:
Screening asymptomatic individuals for early detection of cancer is a public health initiative that is growing rapidly. The main motivation is to diagnose disease early before it progresses to advanced stages so that the benefit of treatment may be enhanced. Models that characterize breast cancer progression are available to evaluate early detection programs. Modeling approaches provide insights into expected outcomes, positive and negative, from early detection programs. Collaborative modeling work of the Cancer Intervention Surveillance Modeling Network (CISNET)-Breast Working Group on mammography screening strategies for the general and high-risk populations will be presented.

Over the past 30 years, malignant melanoma incidence has increased worldwide. Despite this trend, the value of melanoma screening in being debated. Randomized screening trials providing definitive proof of a benefit are not feasible given its prohibitive costs. Sex-specific natural history models for melanoma progression are developed. Using the models, various screening strategies for early detection of melanoma are compared. Basic to any screening programs is when to begin screening, subsequent screening intervals, and if/when to stop screening. Screening schedules with various starting age, stopping age, and screening intervals will be compared systematically to prioritize screening programs, pinpointing those most likely to succeed in reducing mortality and those unlikely to be successful. This is a necessary first step toward ultimately establishing significant/practical guidelines for the general population and/or targeted population having a higher risk of developing melanoma. Our model-projected results on melanoma screening and outcomes will be presented.

Authors/Institutions: S. Lee, Data Sciences, Dana-Farber Cancer Institute, Boston, Massachusetts, UNITED STATES|S. Lee, Harvard Medical School, Boston, Massachusetts, UNITED STATES|S. Lee, Harvard School of Public Health, Boston, Massachusetts, UNITED STATES|
CONTROL ID: 3363931

TITLE: The effect of implementation of organized breast cancer screening and improved treatment on breast cancer mortality the population based screening program in Norway

ABSTRACT BODY:

Abstract Body: Mammographic screening is aimed at detecting breast cancer in an early stage and thus reducing the mortality from the disease. We estimated breast cancer mortality associated with invitations to mammographic screening and the effect of changes in treatment, in Norway.

The population-based breast cancer screening program, offering women aged 50-69 years mammographic screening was implemented stepwise, in the period from 1996 to 2005 in Norway. We used individual-level data for 1,340,333 women from national registries. During 1996–2014 (screening window), women contributed person-years (PY) in non-invited and invited periods. Comparable periods for 1977–1995 (pre-screening window) were created by dividing the follow-up time for each woman into a pseudo-non-invited and pseudo-invited periods. Two different approaches, the regression and matching approaches were used to identify women in the different groups. Breast cancer mortality was estimated for the four periods, using the follow-up and the evaluation model. In the evaluation model we counted breast cancer deaths in each period for all women diagnosed within the period, counting breast cancer deaths and person-years after screening-age only for women diagnosed within screening-age. A multivariable flexible parametric survival model was used to estimate hazard ratio (HR) for the effect of invitation and improved treatment.

Using the regression approach and the evaluation model, we identified 5818 breast cancer deaths and 16,533,281 PY. Invitations to screening reduced BC mortality by 20% (HR: 0.80, 95% CI: 0.70-0.91) among women 50 years and older and by 25% (HR: 0.75, 95% CI: 0.65-0.86) among screening-aged women. The treatment effect was 23% (HR: 0.77, 95% CI: 0.65-0.92) for women 50 years and older, and 17% (HR: 0.83, 95% CI: 0.74-0.94) for screening-aged women. The effect of an invitation was larger among women aged 50-59 versus 60-69, while the effect of treatment was largest among women aged 60-69. Using the follow-up model resulted in somewhat lower, but still comparable with those from the evaluation model.

Implementation of BreastScreen Norway and improved treatment has reduced the mortality substantially during the last decades.

AUTHORS/INSTITUTIONS: S. Hofvind, E. Botteri, S. Sebuødegård, Cancer Registry of Norway, Oslo, NORWAY|S. Hofvind, Oslo Metropolitan University, Oslo, NORWAY|
There has been emerging interest in improving the performance of Phase I (Ph1) dose escalation study design to accurately estimate maximum tolerate dose (MTD), particularly in Oncology. Recently, with the introduction of immunotherapy in oncology, delayed effect has become of great interest to evaluate the toxicity of the agent. This resulted in the need of extending the DLT observation period, but majority of the Ph1 dose escalation studies were designed based on the toxicities observed during the pre-specified dose limiting toxicity (DLT) observation period such as 3 or 4 weeks to avoid much longer study duration. This approach may work if the objective is to identify acute toxicities but it would not be ideal if the agent is expected to have delayed toxicities as it does not properly use the DLT information from post DLT observation period. Therefore, it could potentially result in underestimating overall DLT rate. Here we introduce simple frequentist approach to help dose escalation decision by incorporating DLT observed in both during and post DLT observation period while a conventional DLT observation period can still be used. DLT rate is estimated based on exposure-adjusted event rate manner. The DLT information from earlier cohort can continue to be updated for longer exposure time after DLT observation period, and the information is incorporated in the estimation of DLT rate of the current cohort. Once the dose escalation evaluation is completed, Maximum Tolerate Dose is estimated using isotonic regression. Simulation study will show the performance improvement.

ABSTRACT BODY: The increasing availability of large health care registries has led to an increasing demand for methods that attempt to investigate causal effects on the basis of observational data. Several approaches exist and all of these depend on models of one or more of the processes involved, e.g. models for the outcome or the propensity score. Hence, there is increased interest in methods that are robust against misspecification of the models involved.

In the present project, we pursue a robust method for causal inference, in a setting, where the continuous exposure has missing information and the outcome is binary. When the exposure is continuous, one approach is to use the generalized propensity score (Hirano & Imbens, 2004). We propose to combine this approach with existing methods for dealing with missing data, much like Zhang et al. (2013) proposed for the setting with a binary exposure. Furthermore, when the exposure is continuous one effect measure of interest is the dose-response curve, which for a fixed level of the exposure denotes the expected outcome had the entire target population had that level of exposure. However, we also consider an alternative effect measure, which instead considers the effect of a fixed increase in the exposure from the observed level in the target population. We investigated the performance of our method against other non-robust methods in a simulation study.

Our approach was motivated by an application investigating markers for early disease programming. The objective was to estimate the effect of prenatal vitamin D on the risk of childhood asthma using data from Danish electronic health care registries. The exposure, prenatal vitamin D, is not missing completely at random in the population of interest, since women with low sun exposure or higher socioeconomic status were more frequently measured.

AUTHORS/INSTITUTIONS: M.K. Grand, V. Siersma, Department of Public Health, University of Copenhagen, Copenhagen, DENMARK
Abstract Body: Starting from the observation that the reconstruction of the cohesion among the variates of a multivariate random variable by means of Gaussian graphical model usually takes only sampling variation into account, we point out the consequences of this practice for the reconstruction of the underlying conditional independence graph. When replicates are included in the study, these consequences are overcome by the separation of sampling from other sources of variation. Hereto a simple `signal+noise' model for the description of the multivariate data has been put forward. A penalized EM algorithm for the estimation of the model's parameters has been presented, alongside a discussion of cross-validation for choosing the penalty parameter(s). Through simulation we investigate how much is won by the inclusion of replicates, and compare the presented method to obvious alternatives. Finally, in an illustration using oncogenomics studies with replicates we further investigate the effect of ignoring variation due to other sources than sampling variation and assess the reproducibility of the reconstruction of the conditional independence graph.

Authors/Institutions: W.V. Wieringen, Epidemiology and Biostatistics, Amsterdam UMC, Amsterdam, NETHERLANDS|W.V. Wieringen, Mathematics, VU University Amsterdam, Amsterdam, NETHERLANDS|Y. Chen, Mathematics, Leiden University, Leiden, NETHERLANDS
Abstract Body: A challenging problem in population genetics is to infer the full genealogical history of a sample of DNA sequences—otherwise known as the Ancestral Recombination Graph (ARG)—under the coalescent with recombination. Inferring the ARG remains a problem since, for even a small number of DNA sequences, the state space of the ARG is large.

Many different methods have been proposed to perform the inference, however, most of them have been limited to small datasets. One reason that these methods are not efficient for large sample sizes is because of the way they store and represent the genealogies, i.e., the data structure. Previous methods use a data structure in which each marginal tree is stored separately. This leads to inefficiencies, as neighbouring trees in a genealogy share many parts. In order to gain efficiency and reduce processing time and storage capacity, taking these similarities into account is key.

In 2016, an efficient data structure known as Tree Sequence Recording (TS) was introduced by Kelleher, Etheridge, and McVean to store the genealogical trees at each site. In this method, identical parts of consecutive trees are stored only once. More recently, an inference method—tsinfer—was proposed to infer whole-genome genealogies. This method leverages the features of TS and is applicable to large data sets.

tsfier infers the genealogical trees at each site, however, it is not a probabilistic inference model. Rather, it concentrates on compactly storing large datasets in a novel “evolutionary encoding” format that enables more efficient access and processing of the data.

In this work, we present a Markov chain Monte Carlo (MCMC) approach to perform probabilistic inference under the coalescent with recombination. Borrowing the idea of storing the genealogies with no repeated information from TS, we introduce a data structure to represent the full ARG.

Under the infinite sites mutation model, we infer the full ARG and, unlike tsinfer, our method infers both genealogical trees and event times. Hence, the time to the most common ancestor, the ancestral state at each time, and the total branch length are obtained.

We demonstrate the utility of our method by applying it to simulated datasets. Also, we compare our method with ARGweaver, the state-of-the-art probabilistic method.

Authors/Institutions: A. Mahmoudi, D. Balding, Y. Chan, School of Mathematics and Statistics, The University of Melbourne, Melbourne, Victoria, AUSTRALIA| A. Mahmoudi, D. Balding, Y. Chan, Melbourne Integrative Genomics, The University of Melbourne, Melbourne, Victoria, AUSTRALIA| D. Balding, School of BioSciences, The University of Melbourne, Melbourne, Victoria, AUSTRALIA|
Abstract Body: In life sciences random forests are often used to train predictive models, but it is rather complex to gain any explanatory insight into the mechanics leading to a specific outcome, which impedes the implementation of random forests in the clinical practice. Typically, variable importance measures are used, but they can neither explain how a variable influences the outcome nor find interactions between variables; furthermore, they ignore the tree structure in the forest in total. A completely different approach is to select a single or a set of a few trees from the ensemble which best represent the forest. It is hoped that by simplifying a complex ensemble of decision trees to a set of a few representative trees, it is possible to observe common tree structures, the importance of specific features and variable interactions. The intuitive definition of representative trees are those with the minimal distance to all other trees, which requires a proper definition of the distance between two trees. For this, we developed a new tree-based distance measure and compared it with existing metrics in a simulation study. We show that our new distance metric is superior in depicting the differences in tree structures. Furthermore, we found that the most representative tree selected by our method has the best prediction performance on independent validation data compared to the trees selected by other metrics. Additionally, we observed that in most scenarios a subset of three to five most representative trees gives more accurate predictions than the complete random forest. So we identified the removement of poorly grown trees as a further use case of tree-based distance measures. Finally, we compared all methods on a clinical data set and made them publicly available in an R package.

Authors/Institutions: B. Laabs, I.R. König, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, GERMANY
Complex tissues are composed of a large number of different types of cells, each involved in a multitude of biological processes. Consequently, an important component to understanding such processes is understanding the cell-type composition of the tissues. Estimating cell type composition using high-throughput gene expression data is known as cell-type deconvolution. In this talk, we first summarize the extensive deconvolution literature by identifying a common regression-like approach to deconvolution. We call this approach the Unified Deconvolution-as-Regression (UDAR) framework. While methods that fall under this framework all use a similar model, they fit using data on different scales. Two popular scales for gene expression data are logarithmic and linear. Unfortunately, each of these scales has problems in the UDAR framework. Using log-scale gene expressions proposes a biologically implausible model and using linear-scale gene expressions will lead to statistically inefficient estimators. To overcome these problems, we propose a new approach for cell-type deconvolution that works on a hybrid of the two scales. This new approach is biologically plausible and improves statistical efficiency. We compare the hybrid approach to other methods on simulations as well as a collection of eleven real benchmark datasets. Here, we find the hybrid approach to be accurate and robust.
In Sweden, there are about 90,000 deaths per year, and in about 5,500 of these a forensic autopsy is performed. During these autopsies, samples are collected and submitted to the Department for Forensic Genetics and Forensic Toxicology at the Swedish National Board of Forensic Medicine, where they are routinely screened for several substances, including several pharmaceutical and illicit drugs.

In an accidental death there is interest not only to determine the presence of drugs, but also to determine whether or not it comes from a prescription or an illicit source. This distinction is easy when it comes to illicit drugs, but harder when dealing with pharmaceutical ones.

In this study, we apply multinomial logistic regression to data from both the Swedish National Board of Forensic Medicine and the Swedish Prescribed Drug Register in order to show differences in proportions of licit and illicit use of both pharmaceutical and illicit drugs in different kinds of accidental deaths. We will also illustrate how the use of waiting time distribution can be used or estimate the proportion of illicit use of pharmaceutical drugs.

**AUTHORS/INSTITUTIONS:** S. Gustavsson, A. Jönsson, F. Kugelberg, Department of Forensic Genetics and Forensic Toxicology, Swedish National Board of Forensic Medicine, Linköping, SWEDEN | A. Jönsson, J. Ahlner, F. Kugelberg, Department of Medical and Health Sciences, Linköping University, Linköping, SWEDEN |
Abstract Body: Recent advances in biological research have seen the emergence of high-throughput technologies with numerous applications that allow the study of biological mechanisms at an unprecedented depth and scale. A large amount of genomic data is now distributed through consortia like The Cancer Genome Atlas (TCGA), where specific types of biological information on specific type of tissue or cell are available. In cancer research, the challenge is now to perform integrative analyses of high-dimensional multi-omic data with the goal to better understand genomic processes that correlate with cancer outcomes, e.g. elucidate gene networks that discriminate a specific cancer subgroups (cancer sub-typing) or discovering gene networks that overlap across different cancer types (pan-cancer studies). In this paper, we propose a novel mixed graphical model approach to analyze multi-omic data of different types (continuous, discrete and count) and perform model selection by extending the birth-death MCMC (BDMCMC) algorithm initially proposed by Stephens (2000). We compared the performance of our BDMCMC algorithm to the LASSO method by simulations and found that our method was superior in terms of both computational complexity and model selection results. Finally, an application to the TCGA breast cancer data showed that integrating genomic information at different levels (mutation and expression data) led to better subtyping of breast cancers.

AUTHORS/INSTITUTIONS: L. Briollais, N. Wang, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, CANADA|L. Briollais, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, CANADA|H. Massam, Statistics, York University, Toronto, Ontario, CANADA|
Abstract Body: This talk discusses methods for using a longitudinal biomarker to dynamically predict an interval-censored time to event outcome, using the variability in the longitudinal biomarker. We first investigate a shared random effects model with longitudinal and interval censored survival sub-models. In our motivating clinical example, the biomarker values were highly variable, and the higher the variance likely meant the patient was not adhering to treatment. Thus, individual variance of the longitudinal biomarker was thought to be important in prediction of adverse events. The shared random effects model incorporates the sharing of an individual-specific variance component, along with a traditional intercept and slope. Using this model, we develop a dynamic prediction framework to calculate individualized predicted probabilities of event-free survival for new subjects, based on historical biomarker measurements and demographic data. We show how this model and prediction methodology can be applied clinically, using a dataset from the University of Colorado transplant center. In this example, we demonstrate using a patient’s historical Tacrolimus data, a common immunosuppressant, to predict time-to-adverse kidney graft events.

Authors/Institutions: K.R. Campbell, Pediatrics, University of Colorado, Saint Paul, Minnesota, UNITED STATES| S. Davis, J. Cooper, Division of Renal Diseases and Hypertension, University of Colorado, Denver, Colorado, UNITED STATES| E. Juarez-Colunga, Adult and Child Consortium for Health Outcomes and Delivery Science, University of Colorado, Denver, Colorado, UNITED STATES| E. Juarez-Colunga, G.K. Grunwald, Biostatistics and Informatics, University of Colorado, Denver, Colorado, UNITED STATES|
CONTROL ID: 3366414

TITLE: The association of exposure to and engagement with tobacco-related social media content and the age of initiation of e-cigarettes among USA youths in 2013-2017

ABSTRACT BODY:

Abstract Body: In the USA, there are no federal laws against tobacco advertising on social media sites. In 2019, Facebook, Instagram, and Snapchat updated their policies to prohibit brand advertisements and user-generated content for the sale of tobacco and e-cigarettes. Tobacco users and influencers can still upload their own posts, photos, or videos of themselves using and promoting these products. Social media sites have the potential to expose youth to tobacco-related content as 51-72% of the USA teens ages 13 to 17 used these platforms in 2018. Analyses for interval-censored data from 2013-2017 of the Population Assessment of Tobacco and Health youth dataset (ages 12-17; n = 12,701; N = 23,313,039) were carried out. Age of initiation was prospectively estimated among non-users of e-cigarettes using participant age in 2013, calendar week/year of survey participation, and the number of weeks between the last report of never use and the first report of ever e-cigarette use. Survival functions overall and by sex are reported. In 2013, two potential risk factors for e-cigarette initiation were assessed: (i) “In the past 12 months, have you seen any tobacco-related content on social media sites?”, and (ii) “In the past 12 months have you posted content about tobacco products on any social media sites?”. Hazard ratios (HR) and 95% CI were estimated, adjusting for the interaction of sex and race but there were not meaningful differences. Overall, 24% (N = 5,647,835) of youth initiated e-cigarettes between 2014 and 2017. Cumulatively, we found that 13% of youth initiated e-cigarettes by age 15, 20% by age 16, 32% by age 17, 44% by age 18, and 49% by age 20. The risk of initiating e-cigarettes is 1.536 times higher in participants who had seen tobacco-related content on social media as compared to participants who had not seen tobacco-related content on social media (95% CI: 1.533-1.539). The risk of initiating e-cigarettes was 1.782 times higher in participants who have posted tobacco-related content on social media as compared to those who did not (95% CI: 1.776-1.788). Exposure to and engagement with social media content specific to tobacco and other e-cigarette use increases the risk of initiating e-cigarette use among youth. Stronger policy-based solutions are needed to reduce tobacco-related content on social media.

AUTHORS/INSTITUTIONS: A. Perez, B. Chen, Biostatistics and Data Science, The University of Texas Health Science Center at Houston, Austin, Texas, UNITED STATES|A. Perez, M. Bluestein, B. Chen, C. Perry, M. Harrell, Michael & Susan Dell Center for Healthy Living, UTHHealth, Austin, Texas, UNITED STATES|M. Harrell, Epidemiology, The University of Texas Health Science Center at Houston, Austin, Texas, UNITED STATES|
Abstract Body: With modern technology development, more and more diagnostic markers are being collected as functional data (i.e., functional markers). Each sample element of a functional marker is a smooth, continuous curve, whose dynamic structure over a time or space domain is a rich source of clinical information. In many clinical practices, it is standard to describe and diagnose a disease using a set of "quantitative features" that characterizes various dynamic, interpretable patterns of a functional marker, such as area under the curve, maximum value and time to reach maximum. Here, we present a novel statistical framework for evaluating the diagnostic accuracy of quantitative features using the area under the receiver operating characteristic curve (AUC). Based on a class of summary functionals that flexibly represents various quantitative features, we develop a two-stage non-parametric AUC estimator that addresses discreteness and noise in functional data and establish its asymptotic properties. To describe the heterogeneity of AUC in different subpopulations, we propose a sensible adaptation of a semi-parametric regression model, whose parameters can be estimated by the proposed estimating equations. We also propose an automated data-driven approach for trading off between bias and efficiency of the regression coefficient estimates when continuous covariates are considered. We demonstrate the application of our methods using a renal study.

AUTHORS/INSTITUTIONS: J. Jang, Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, UNITED STATES|A. Manatunga, Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, UNITED STATES|
Abstract Body: The National Health and Nutrition Examination Survey (NHANES) is a major program of the National Center for Health Statistics, designed to assess the health and nutritional status of adults and children in the United States. The analysis of NHANES dental caries data faces several challenges, including (1) the data were collected using a complex, multistage, stratified, unequal-probability sampling design; (2) the sample size of some primary sampling units (PSU), e.g., counties, is very small; (3) the measures of dental caries have complicated structure and correlation, and (4) there is a substantial percentage of nonresponses, which are expected to be not missing at random or non-ignorable. We propose a Bayesian hierarchical spatial model to address these analysis challenges. We develop a two-level Potts model that closely resembles the caries evolution process and captures complicated spatial correlations between teeth and surfaces of the teeth. By adding Bayesian hierarchies to the Potts model, we account for the multistage survey sampling design and also enable information borrowing across PSUs for small area estimation. We incorporate sampling weights by including them as a covariate in the model and adopt flexible B-splines to achieve robust inference. We account for non-ignorable missing outcomes and covariates using the selection model. We use data augmentation coupled with the noisy Monte Carlo method to overcome the numerical difficulty caused by doubly-intractable normalizing constants and sample posteriors. Our analysis results show strong spatial associations between teeth and tooth surfaces and that dental hygienic factors, fluorosis and sealant reduce the risks of having dental diseases. This is the joint work with F. Liu, E. Eugenio, K. You and S. Liu.

AUTHORS/INSTITUTIONS: I. Jin, Department of Applied Statistics, Yonsei University, Seoul, KOREA (THE REPUBLIC OF)|F. Liu, ACMS, University of Notre Dame, Notre Dame, Indiana, UNITED STATES|S. Liu, Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, UNITED STATES|
Abstract Body: Tomography aims to display cross-sections through human bodies or other solid objects using data collected from around the body or object. Two main branches are positron emission tomography and electrical impedance tomography. Modern positron emission tomography datasets record about 1 billion radioactive events as more than 10,000 highly correlated but separate counts. A common aim is to estimate parameters and represent the spatial distribution of radioactive tracer concentration. This makes such applications challenging inverse problems.

From a statistical perspective, inverse problems are regression models where a response, depending on a number of causal factors or parameters, is measured and the goal is to estimate the parameter values. Since in typical tomography applications inverse problems may be highly multivariate and have predictors which are highly correlated, even simple linear problems cannot be solved by classical regression methods, nor can they be adequately solved using standard dimension reduction or regularised regression techniques. A remedy is to use Bayesian hierarchical models, which can be slow to fit via standard MCMC algorithms.

We propose a mean field variational Bayes (MFVB) approach for accurate approximate fitting and inference on inverse problem models. In parallel, we identify typical factor graph fragments arising in inverse problem Bayesian models and derive the variational message passing version of the MFVB algorithms. The resultant factor graph fragments facilitate streamlined implementation of approximate algorithms for inverse problems motivated by tomography and set the basis for software development. As a matter of fact, the factor graph fragment paradigm allows easy incorporation of different penalization structures in the model or changes to the distribution of the outcome variable. Nevertheless, variational message passing on factor graph fragments is such that algorithm updates and streamlining steps only need to be derived once for a particular fragment and can be used for any arbitrarily complex model including such fragment. Hence dramatically reducing set-up overheads as well as providing fast implementation.

AUTHORS/INSTITUTIONS: L. Maestrini, M.P. Wand, Faculty of Science, School of Mathematical and Physical Sciences, University of Technology Sydney, Sydney, New South Wales, AUSTRALIA|R.G. Aykroyd, University of Leeds, Leeds, UNITED KINGDOM|L. Maestrini, M.P. Wand, Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers, Melbourne, Victoria, AUSTRALIA|
Nigeria is the most populous country in Africa and the biggest economy, and has been experiencing the most violent insurgency in the country’s history. Therefore, it is very crucial to understand the spatial and temporal distribution of insurgent activities to improve knowledge about attacks patterns, planning and counterterrorism measures. Using geostatistical analysis of incidents tracking system database, we provide insight into the insurgent incidents in Nigeria from 2000 to 2017. This study investigates the pattern of violent insurgency rates and explores the social, political and religious triggers that influence these incidents. The analysis in this study include incident density map, identifying the dominant socio-ethnic-religion group where incidents occurred, investigating the spatial and spatio-temporal pattern of incidents.

AUTHORS/INSTITUTIONS: O.A. Adegboyé, Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam, Ho Chi Minh, VIET NAM|F.M. Elfaki, Qatar University, Doha, QATAR|M.A. Adegboyé, O. Jamiu, American University of Nigeria, Yola, NIGERIA|T. Saffary, Independent Research, MD, Maryland, UNITED STATES|
Abstract Body: In cancer clinical trials, the sum of the longest diameter (SLD) of target lesions is a biomarker which reflects both the tumor burden and its evolution over time. Modelling this biomarker jointly with survival times improves the inference about treatment and prediction accuracy of the survival time. An excess of zero values and right skewness often characterize the SLD distribution. While a nonlinear transformation can easily handle the latter, the zero-inflation problem requires a more sophisticated approach. Left-censoring has been proposed (Król, A. et al. Biometrics 2016) as a way to handle the excess of zero values in a mixed-effects model (i.e. values below a detection limit are censored). Patients responding well to a treatment can reach the complete response state as defined by RECIST criteria, in which case the tumor size shrinks until reaching a ‘true zero’ value (i.e. not censored). We propose a two-part joint model that decomposes the distribution of the biomarker into a binary outcome (zero values vs. positive values) and a continuous outcome, both outcomes being modelled by a mixed effects regression model. Therefore, the binary part captures the effect of covariates on the probability of zero value of the biomarker. We propose two forms for the continuous part: the conditional form captures the effect of covariates on the expected value of the biomarker among positive values and the marginal form captures the effect of covariates on the marginal mean of the biomarker (Smith, V. A. et al. Stat. Med. 2014). The survival times are modelled using a Cox proportional hazards model with splines approximation for the baseline hazard. We propose several association structures to link the biomarker to the risk of event. We illustrate with simulation studies the performances of the models in terms of bias, accuracy and coverage probabilities when the model assumptions are misspecified, with different rates of zero excess. A real data application to a colorectal metastatic cancer trial comparing two treatment strategies is also proposed, showing how the biomarker evolution over time can bring additional information to evaluate the treatment effect on the risk of death. The zero inflation is a common problem in biomedical research, e.g. when quantifying exposure or measuring symptoms of a disease, and our proposed model is also relevant in this wide spectrum of applications.

Authors/Institutions: D. Rustand, V. Rondeau, Bordeaux Population Health Center, Bordeaux, FRANCE|L. Briollais, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, CANADA|L. Briollais, Biostatistics, Dalla Lana School of Public Health, Toronto, Ontario, CANADA
Abstract Body: Clinical risk prediction models aim to predict the presence (diagnostic) or future risk (prognostic) of disease, and there exists a growing body of literature on the development and evaluation of such models in many disease areas. Intuitively, it is desirable that prediction models include causal risk factors. However, despite recent progress in causal inference, incorporating their results in clinical risk prediction models has remained largely unexplored. Here, we assess the usefulness of two causal inference concepts based on Directed Acyclic Graphs (DAGs) for clinical risk prediction modeling, theoretically and based on a Monte Carlo simulation study.

First, we investigated whether incorporating knowledge about the underlying causal structure can provide insights on the transportability of diagnostic clinical risk prediction models, where transportability refers to the application of the model in a population with a different distribution of the predictor. Using two example scenarios, we demonstrate that this is indeed the case and that a prediction model that correctly assumed the causal direction of predictor and outcome had better transportability than one that assumed the anticausal direction.

Second, we asked whether causal knowledge can be used to a priori select the optimal set of predictors for an outcome. We randomly generated 100,000 DAGs and compared a logistic regression model including the Markov Blanket set – which are the parents, children, and parents of the children of the outcome node – in a DAG as covariates with seven other methods including three logistic regression models using other criteria to select the predictors, lasso, ridge, and elastic net regression models, and a random forest algorithm. The results indicated that the Markov Blanket is the optimal set of predictors for the outcome, and yielded risk prediction models with best predictive performance. Hence, using DAGs to identify Markov Blanket variables may be a useful and efficient strategy to select predictors in clinical risk prediction models if strong knowledge of the underlying causal structure exists or can be learned.

Taken together, these two results strongly support to incorporate knowledge about the underlying causal structure as well as recent methodological advances from causal inference into clinical risk prediction modeling.
Abstract Body: We introduce a novel topological data analysis (TDA) framework for analyzing the pattern of trees (graphs with no loops). Due to the lack of one-to-one correspondence between trees at the node or edge level, building a coherent statistical model has been a challenge. It is not even clear how to set up a simple linear model across trees when there is no correspondence. To address the problem, we present an algebraic representation called the Weighted Fourier Series (WFS) that was originally developed for parametrically representing functional data on manifolds (Statistica Sinica 18:1269-1291). WFS can be used to represent trees using the eigenfunctions of the Laplace-Beltrami operator of the underlying manifold with almost no reconstruction error. WFS enables us to establish the one-to-one correspondence between trees by establishing the diffeomorphic point correspondence in the underlying manifold. Statistical models can be easily set up across the expansion coefficients of the representation or on the representation itself. The method can be also used to define the inner product between trees in the proper Hilbert space without any complicated tree operations often done in the tree space methods.

We applied the method to the 3-Tesla magnetic resonance images (MRI) dataset consisting of 268 females and 176 males collected as the subset of the Human Connectome Project (arXiv 1911.02721). The cortical surfaces of the brain were segmented and used in further extracting sulcal and gyral curves that are grooves and ridges in the cerebral cortex. The sulcal and gyral curves are shaped like a collection of disconnected trees and characterize the overall shape of the cerebral cortex. The curves are then mapped to a sphere and spherical harmonics are then used as a basis in WFS. Persistent homology techniques are subsequently applied to WFS of sulcal and gyral curves in quantifying the both local and global topological differences between the groups. We used the persistent diagrams as the topological summary of sulcal and gyral curves. For the statistical inference, the permutation test in the spectral domain was used to preserve the topological invariance under the null hypothesis of no group difference (Annals of Applied Statistics 12:1506-1534).

AUTHORS/INSTITUTIONS: M.K. Chung, Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, UNITED STATES
Abstract Body: Bivariate recurrent event data are observed when a subject has experienced two different type of events repeatedly. For analyzing such multivariate recurrent event time data, bivariate frailty effects have been applied to reflect both the association between events and one within each event. In this paper, our interest is to suggest statistical model when there is a substantial portion of subjects not experiencing recurrent events but having a terminal event. In a context of recurrent event, zero events can occur either at the cure group which is unsusceptible to events or at the censoring before the occurrence of recurrent events. For simultaneously reflecting both a zero inflation and a terminal event, a joint model is implemented with bivariate frailty effects. In particular, two joint models are considered depending on the homogeneity of death hazards among susceptible groups. Simulation studies are performed to evaluate the suggested models. Infection data from AML (acute myeloid leukemia) patients are analyzed as an application.

Authors/Institutions: Y. Kim, Statistics, Sookmyung Women's University, Seoul, KOREA (THE REPUBLIC OF)
Abstract: A common aspect of count data is that the observed frequency of zero counts is unusually large, typically with respect to the Poisson distribution. Specific modelling approaches include both overdispersion and zero-inflation models. Although these are sometimes seen as competing approaches, they can be complimentary: overdispersion is a consequence of zero-inflation models and, in the same way, zero-inflation is a by-product of overdispersion models. In general the common focus of any analyses is on inference for the effect of covariates on the mean, however, in view of the excess zero counts we may also be interested in modelling this aspect. This paper will review the different approaches to zero-inflation modelling, in particular where this relates to forms of zero dependence that can only be generated through explicit modelling of the zero-inflation. The aim is to be able to disentangle zero-inflation and overdispersion covariate dependence as functions of the zero probability of the base count distribution (typically the Poisson). To aid in this we propose diagnostic plots that may be helpful in making critical evaluations of different modelling approaches. Several examples will be used to illustrate these ideas.

Authors/Institutions: J. Hinde, NUI Galway, Oughterard, Galway, IRELAND|J. Haslett, Trinity College Dublin, Dublin, IRELAND|A. Parnell, Maynooth University, Maynooth, IRELAND
Abstract Body: Summarizing multiple correlated omics datasets for dimension reduction and for interpretation is an open research topic. Various methods have been proposed for this purpose, such as PLS-related approaches, which decompose datasets into joint and residual parts. Omics data are heterogeneous (e.g. differences in source of variation, scale, dimensionality, etc.) and the joint parts estimated in PLS contain data-specific variations. O2PLS was proposed to capture the heterogeneity using data specific parts and better estimate the joint parts. However, the latent components spanning the joint subspace in O2PLS are linear combinations of all the variables, hampering interpretation. For better interpretation, variable selection is needed. To this end, we extend O2PLS to Group Sparse O2PLS (GSpO2PLS) which performs variable selection and incorporate group structures of variables.

Our motivating datasets are methylation (482,563 CpG sites) and IgG glycomics (22 glycan peaks) data from 646 samples in the TwinsUK study. IgG is an antibody whose functional diversity is mainly achieved by glycosylation. Methylation has an important role in the glycosylation pathways. We aim to identify groups of CpG sites that impact IgG glycosylation, and hence have influence on immune response.

To perform variable selection, L1 penalty is introduced on the loadings. Sparse solutions are obtained by retaining only variables with a large contribution to the covariance. GSpO2PLS imposes penalties on the sum of the group-wise L2 norms of loadings, which result in group-wise sparsity where variables of the same group are selected or dropped altogether based on the contribution to the covariance as a whole. If all the groups have size 1, the sum of group-wise L2 norms corresponds to L1 norm.

A simulation study shows that GSpO2PLS performs better than O2PLS in terms of variable selection and prediction, especially when group information is available. We apply GSpO2PLS to the motivating datasets, where we take CpG sites located around the same gene as a group. The genes corresponding to the selected regions of CpG sites are crucial for protein glycosylation and immune functions.

GSpO2PLS provides a framework to integrate two heterogeneous omics datasets and select relevant groups of variables, thereby facilitating interpretation.

Authors/Institutions: Z. Gu, S. El Bouhaddani, H. Uh, Biostatistics & Research Support, University Medical Center Utrecht, Utrecht, NETHERLANDS|J. Houwing-Duistermaat, Department of Statistics, University of Leeds, Leeds, UNITED KINGDOM
Abstract Body: Understanding the temporal interaction amongst species within an ecological community has been a critical challenge in biodiversity research. The species interaction often exhibits a feedback mechanism, in which the outputs of the past can be the inputs to the current state of ecological systems. Given the observations of multi-species time series, we developed a time series model for a multivariate feedback system, adopting linear interaction structures between species as a tractable approximation. The feedback model is then examined in the spectrum domain to delineate the extent to which a given species is driven by other species at different frequency ranges, in other words, longer or shorter time lag effects as interactions. Our results show that the modelling approach offers new insights into temporal species interaction in an ecological community. We demonstrate how it enables further analysis into ecologically relevant groups of species that underpin the dynamics of the system.

Authors/Institutions: H. Shimadzu, Mathematical Sciences, Loughborough University, Loughborough, Leicestershire, UNITED KINGDOM|H. Shimadzu, Graduate School of Public Health, Teikyo University, Tokyo, JAPAN|
Exploring heterogeneous treatment effects using machine learning to inform the targeting of national health insurance programmes

Researchers evaluating the effects of social policies, such as health insurance, are interested in identifying subgroups who would benefit the most, to inform the design of expansion programmes when resources are limited. However, many policy evaluations stop at estimating average treatments effects, potentially masking health benefits on subpopulations.

In this study, we explore treatment effect heterogeneity and estimate conditional average treatment effects (CATE) of being enrolled in social health insurance scheme in the year of childbirth on having an assisted birth, in a retrospective cohort of Indonesian mothers, between 2002 and 2014. The CATEs are estimated for both, pre-specified (receipt of cash transfer as a proxy for poverty) and data-driven effect-modifiers.

We perform two approaches that incorporate machine learning, assuming that we have a sufficient set of variables for confounding control and that the positivity assumption holds. To deal with missing values, we use inverse probability of being a complete case weights.

For pre-specified subgroups, we use T-learner and X-learner approaches, which use estimates of the individual-level predictions of both potential outcomes to calculate individual-level treatment effects. Here, we implement these with parametric (main terms logistic) and random forests (RFs) models. We then use Causal Forests, which train the RFs directly on predicted individual-level treatment effects which are estimated in a first stage, to obtain data-driven effect modifiers and estimate the corresponding CATEs.

For the pre-specified subgroup, all methods resulted in similar inferences, namely there is no evidence that receipt of cash transfers is an effect modifier. However, Causal Forest test of heterogeneity is significant. We find monotonously decreasing relationships between treatment effects and education levels and household wealth.

These findings indicate that the social health insurance programmes would benefit the most those from more disadvantaged backgrounds, and that pre-specified subgroups are not always the place where we should be focusing. Given major fiscal constraints, studying treatment effects heterogeneity in a data-adaptive fashion can help defining an “optimal treatment rule” for eligibility of free health insurance and could help improve the efficient allocation of resources.

K. DiazOrdaz, Medical Statistics, LSHTM, London, UNITED KINGDOM|N. Kreif, Centre for Health Economics, University of York, York, UNITED KINGDOM|
Identifying the most parsimonious model in Structural Equation Modelling is of utmost importance and the appropriate power estimation methods minimize the probabilities of Type I and Type II errors. The power of the test depends on the sample size, Type I error, Degrees of Freedom and Effect Size. In this study, a modified approach of calculating the non-centrality parameter for power is proposed. This is compared to MacCallum et al. (1996) at different sample size specifications (from 50 to 2000). For smaller sample sizes and degrees of freedom, the power of the tests for the modified approach were smaller than the traditional. However, as the sample size increased, the difference between the power of the test for both methods reduced to zero. The results showed that the values for power of the test are the same for modified and the traditional approach under large sample sizes and degrees of freedom.

AUTHORS/INSTITUTIONS: A.O. ADEBANJI, S. Appiah, DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, Ashanti, Ashanti, GHANA|D. Adedia, Department of Basic Sciences, University of Health and Allied Sciences, Ho, Volta, Ghana|
Abstract Body: Background: Predicting frequent emergency department (ED) use helps healthcare providers adopt appropriate interventions to reduce the cost and increase the quality of care. This study aims to assess and compare the predictive ability of machine learning (ML) models in identifying frequent ED users.

Method: Korean Health Panel data from 2008 to 2015 were analyzed for this study. Individuals with at least one ED visit were included, among whom those with four or more visits per year were considered to be frequent users. Patients’ demographic and ED-related predictors were included in the analysis. Logistic regression (LR), random forest (RF), SVM, as well as two ensemble models, namely bagging and voting, were trained and tested to examine their predictive performance.

Results: The strong predictors of ED frequent visits included age over 65 years, day of visit (e.g. Sunday), reason for visit (e.g. respiratory system), transportation (e.g. own car), and insurance type (e.g. national health insurance). All the predictive techniques identified frequent ED users with high precision (98%) and sensitivity (RF=94%, the highest; LR=74%, the lowest). ML algorithms showed high AUC values (92%) but good to poor classification error; LR had the highest classification error (14.7%) while RF had the least (4.1%).

Conclusions: The results show that predictor variables in administrative health panels are reliable indicators for predicting frequent ED use. The performance of ML algorithms also shows a promising path in predicting future events using population health data.

Authors/Institutions: H.J. Lim, R. Safaripour, Department of Community Health and Epidemiology, College of Medicine, University of Saskatchewan, Saskatoon, Saskatchewan, CANADA
Abstract Body: In quantifying the probability of survival in cancer patients using cancer registry data, it is common to estimate marginal relative survival, which under assumptions can be interpreted as marginal net survival. Net survival is a hypothetical construct giving the probability of being alive if it was only possible to die of the cancer under study and enables comparisons between populations with differential mortality rates due to other causes. Cause of death information is either unavailable or, more commonly, deemed unreliable and so expected mortality rates are used so that mortality in excess of that expected can be estimated.

Marginal relative survival can be estimated non-parametrically (Pohar Perme estimator) or in a modelling framework. In the modelling framework, even when just interested in the marginal survival in a population it is necessary to model covariates that affect the expected mortality rates (e.g. age, sex and calendar year). The marginal relative survival function is then obtained through regression standardization. Given that these covariates will generally have non-proportional effects, the model can become complex even before other exposure variables are considered.

We propose a flexible parametric model incorporating restricted cubic splines that directly estimates marginal relative survival and thus removes the need to model covariates that affect the expected mortality rates. In order to do this the likelihood needs to incorporate the marginal expected mortality rates at each event time taking account of informative censoring. In addition time-dependent weights need to be incorporated into the likelihood. An approximation is proposed using data expansion which enables the marginal model to be fitted using standard software to fit relative survival models.

The advantage of this approach is that when there is interest in marginal effects there is no need to model the effect of covariates for which the expected mortality rates vary and thus simplifies the modelling process. Additional weights can be incorporated when needing to standardize to an external reference population, which is common in these studies.

The methods are illustrated using English cancer registration data. In addition a simulation study has been performed comparing estimation using a non-parametric approach, regression-standardization and the new marginal model showing that the method has good statistical properties.

AUTHORS/INSTITUTIONS: P. Lambert, E. Syriopoulou, M.J. Rutherford, Health Sciences, University of Leicester, Leicester, UNITED KINGDOM| P. Lambert, Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, SWEDEN
Abstract:

Background: Available parameter estimation methods for the joint modelling of time-to-event and longitudinal outcomes have typically only allowed for a single event outcome and a single longitudinal outcome. In many clinical trials, patients who are followed up over time may often experience a succession of multiple events and further patient's health status may be followed up by several quantitative longitudinal biomarkers. In this context, we presented a joint multistate Markov model for predicting the clinical progression of HIV infection which takes into account the viral load count biomarker.

Method: It is implemented under SAS®, by using both SAS PROC NLMIXED and %JM SAS Macro programming. The SAS PROC NLMIXED and %JM SAS macro programs allows researchers to estimate the parameters of this joint multistate Markov model within the maximum likelihood framework using a quasi-Newton algorithm.

Results: The model confirmed that viral load dynamics significantly affect the transition intensities of HIV/AIDS disease progression. Patients with many sexual partners and with high liver enzymes abnormality showed significantly reduced intensities of immunological recovery transitions. Furthermore, a high weight, high education levels, high QoL scores, high RBC parameters and being of middle age significantly increased the intensities of immunological recovery transitions.

Conclusions: A joint multistate Markov model approach provides wide-ranging information about the progression and assists to provide specific dynamic predictions and increasingly precise knowledge of diseases.

AUTHORS/INSTITUTIONS: Z.G. Dessie, Statistics, University of KwaZulu-Natal, Durban, KwaZulu-Natal, SOUTH AFRICA|Z.G. Dessie, Statistics, Bahir Dar University, Bahir Dar, ETHIOPIA
Abstract Body: Estimating the effect of treatment strategies from longitudinal observational data often requires g-methods, such as inverse probability (IP) weighting or the iterative conditional expectation (ICE) g-formula. These estimators are singly robust in the sense that there is only one opportunity to get valid estimates. IP weighted estimators require a sequence of correctly specified models for the conditional probability of treatment, and ICE requires a sequence of correctly specified models for the conditional mean of the counterfactual outcome. Multiply robust estimators that combine ICE and IP weighting require that only one of the sequences of models is correctly specified, and thus offers more than one opportunity for valid estimation. This is important in practice because some degree of model misspecification is almost always expected when we use parametric models in our analyses. These multiply robust estimators are also consistent under other model misspecification scenarios: for example, the doubly robust estimator from Bang and Robins (2005) would result in valid estimates if the models for the outcome are correctly specified after a certain follow-up time and the models for the treatment process are correctly specified before that follow-up time.

Though several multiply robust estimators exist, they have never been compared to singly robust methods within the context of survival analysis and so, it is unclear whether the increased technical complexity of these methods is worthwhile. Here, we compare singly robust estimators with various multiply robust estimators for survival outcomes. These multiply robust estimators can be shown to be algebraically equivalent to the augmented inverse probability weighted estimator and thus will be locally asymptotically efficient when all of the models are correctly specified. Via simulations studies we show that multiply robust estimators confer more protection against model misspecification than singly robust estimators, and that certain multiply robust estimators offer more protection than others. Finally, we will compare these methods in an analysis of a large epidemiological study and provide guidelines for practitioners interested in implementing these methods in estimating the treatment effect in survival analyses.

AUTHORS/INSTITUTIONS: L. Wen, M. Hernan, J. Robins, A. Rotnitzky, Harvard University, Boston, Massachusetts, UNITED STATES| A. Rotnitzky, Universidad Torcuato Di Tella, Buenos Aires, ARGENTINA|
Abstract Body: For heterogeneous diseases such as cancer, it is highly unlikely to discover a single biomarker that can be used for all subjects. Constructing a panel of biomarkers may not be useful neither because each biomarker performs differently varying from person to person. An alternative is a personalized diagnostic strategy that applies the more accurate biomarker to each subject according to their individual characteristics. In this paper, we propose a linear combinations of the individual characteristics to identify the subgroup of subjects for inferring optimal and personalized biomarker selections. An efficient grid rotation algorithm that guarantees to increase the area or partial area under the receiver operating characteristic curve compared to the standard one-size-fits-all diagnostic strategy.

Authors/Institutions: Y. Chung, College of Health Solutions, Arizona State University, Phoenix, Arizona, United States|Y. Zhang, School of Mathematics and Statistical Sciences, Arizona State University, Tempe, Arizona, United States|Y. Chung, Biodesign Center for Personalized Diagnostic, Arizona State University, Tempe, Arizona, United States
Abstract Body:
Digital phenotyping studies involve collecting data from an individual's smartphone (such as GPS, call/text logs, surveys) to provide information relevant to psychiatric disorders and other illnesses. These data frequently contain too many explanatory variables, and principal component analysis (PCA) proves to be a useful dimension reduction tool. However, missingness in digital phenotyping data leads to problems with PCA and other downstream analyses. Existing methods for PCA in the presence of missingness rely on imputation, but can have trouble with reduced variance in the features as well as with interpretation of the principal components. We introduce a non-imputation based approach contingent on the pairwise empirical correlation matrix. We show in simulations how our method performs compared to imputation methods under varying correlation structures and proportions of missingness. We also compare competing methods using a digital phenotyping study of schizophrenia.

AUTHORS/INSTITUTIONS: E. Getzen, I. Barnett, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES
Abstract Body: An adverse drug event (ADE) refers to injury resulting from the use of a drug and may be due to a medication error or an adverse drug reaction. ADEs are a persistent and important global problem in hospitals. ADEs occur often, cause high morbidity and increased length of stay, and are costly. Hospitals typically rely on staff to detect ADEs or potential ADEs and to document these via incident reporting systems – resulting in under-reporting. With the increased availability of large electronic health records (EHRs) databases comes the chance of enhancing early detection of ADEs. EHRs contain information, such as a patients’ medical history, diagnoses, medications, and laboratory test results. These huge datasets are fundamentally different from existing administrative summary datasets where much data are collected routinely after patient discharge from hospitals. These new data are sequential and time-stamped to record health care delivery activities. This project demonstrated how different statistical analyses were applied to utilise these dynamic data to detect ADEs with different clinical mechanisms, drawing on two case studies. More than 5.6 million laboratory test results and 2.7 million medication records for 46,000 patient admissions were extracted from different EHR systems. Hyperkalaemia (high potassium) was examined in the first case study. It is one of the common fluid and electrolyte abnormalities with potentially serious consequences. Patients’ potassium levels and medication data were linked based on time sequences. Multilevel multivariable logistic models were applied to identify medication induced hyperkalaemia. The second case study involved acute kidney injury (AKI), a significantly under diagnosed condition which could be fatal and requires intensive treatment. However, AKI may be reversible if diagnosed and treated. An algorithm based on creatinine changes over time was developed to identify patients’ AKI status and stages. Medications with potential to cause AKI were linked and modelled to further identify medication induced AKIs. Both case studies provided insights into the building of statistical real-time models into EHRs to support clinical decision-making to improve patient safety and outcomes.

Authors/Institutions: L. Li, K. Rathnayake, J. Westbrook, Australian Institute of Health Innovation, Macquarie University, Sydney, New South Wales, AUSTRALIA
Abstract Body: In meta-analysis, hypothesis testing is one of the commonly used approaches for assessing whether heterogeneity exists in effects between studies. The literature concluded that the Q-statistic is clearly the best choice and criticized the performance of the likelihood ratio test in terms of the type I error control and power. However, all the criticism for the likelihood ratio test is based on the use of a mixture of two chi-square distributions with 0 and 1 degrees of freedom, which is justified only asymptotically. In this study, we develop a novel method to derive the finite sample distribution of the likelihood ratio test and restricted likelihood ratio test for testing the zero variance component in the random effects model for meta-analysis. We also extend this result to the heterogeneity test when meta-regression is applied. A numerical study shows that the proposed statistics have superior performance to the Q-statistic, especially when the number of studies collected for meta-analysis is small to moderate.
The best way to handle dependence across features when testing a high dimensional parameter has raised many discussions with unclear final recommendations. The former global testing issue arises in a wide scope of applications, including functional Analysis of Variance (fANOVA) and association tests between a region of the genome formed by contiguous Single Nucleotide Polymorphisms (SNP) and a case/control response variable in Genome Wide Association Studies (GWAS). Interestingly, in the two former fields of applications, many popular methods are just based on simple aggregation of pointwise test statistics ignoring their dependence.

Addressing the dependence issue often consists in observing its detrimental impact on the performance of standard methods designed to be optimal under independence, and deduce ad-hoc improvements. To be valid for arbitrarily complex dependence patterns, such approaches can lead to poorly powerful procedures. Therefore, a new generation of methods have emerged, advocating for an ad-hoc handling of dependence based on a preliminary whitening of the data.

After a general discussion on the performance of testing methods ignoring dependence or whitening the pointwise test statistics, we show that none of those two extreme choices is uniformly powerful over the variety of dependence and association patterns. A new class of aggregation methods is therefore introduced, spanning the range between total ignorance of dependence and complete decorrelation. Its performance is discussed both in fANOVA for large Event-Related Potentials (evoked ElectroEncephaloGrams) designs and in SNPset approaches of GWAS.


AUTHORS/INSTITUTIONS:  D. Causeur, Univ Rennes, Agrocampus Ouest, CNRS, IRMAR - UMR 6625, Rennes, FRANCE
Abstract Body: Personalized medicine is a methodology that involves estimation of decision rules that enable choosing one treatment over another treatment in order to improve the chance of a successful treatment outcome.

In personalized nutrition, the same line of thinking may be useful even though, in contrast to the clinical setting in medicine, there is typically a gradient of improvement in a health outcome (e.g., weight loss) when choosing between two diets. However, so far the focus in personalized nutrition has been mostly on individualized diets (e.g., individual dietary advice) rather than on individualized effects of diets.

Although it sounds contradictory, individualized effects of diets may be studied much more efficiently than individualized diets, by means of data from randomized controlled trials (RCTs) with well-defined diets as treatments. It may be by means of tailored subgroup analyses (Hjorth et al., 2017) or by exploiting bio-marker-diet interactions (Ritz et al., 2019). The talk will briefly revisit these two types of approaches for investigating individualized effects of diets.

The biomarker-diet interaction model is a parametric statistical model, providing useful interpretations. However, assuming a parametric model inevitably entails several assumptions. One key assumption is linearity in the biomarker, which may be plausible in certain ranges but may be difficult to justify on biological or physiological grounds. Another limitation is that the prediction error was only characterized using the confidence bands on the individualized treatment effect. The predictive capability of such statistical models could, however, be further explored through evaluation of its usefulness for predicting individualized effects on external data or through cross validation. These extensions will be discussed in the talk.

References


Authors/Institutions: C. Ritz, M. Fiil Hjorth, University of Copenhagen, Frederiksberg C, DENMARK
In randomized clinical trials, methods of generalized pairwise comparisons (GPC) have recently gained much attention when interests lie in assessing the effect of a treatment as compared to a standard of care. Among other advantages, these methods are usually praised for delivering a treatment measure that can easily handle multiple outcomes of different nature while keeping a meaningful interpretation for patients and clinicians. In this project, we focus our attention to the study of time-to-event outcomes in settings where we know that a fraction of the individuals will never experience the event of interest. A common challenge in the presence of a so-called cure fraction is that, when confronted to censoring, one does not observe whether a censored observation is in fact cured or not. To the best of our knowledge, GPC has not been studied in this context.

Based on the well-studied mixture cure model, we start by uncovering the impact of the cure fraction on general methods of pairwise comparisons. In particular, we illustrate how the neglecting of this characteristic of the data may in this case seriously affect the conclusions about treatment effect, as both curative and life-prolonging features are then encompassed in one single treatment measure that loses its attractiveness in terms of interpretability. As a remedy, we then suggest a simple estimation procedure that aims to disjoint the curative effects of a treatment from its life-prolonging ones for pairwise comparison methods. The resulting methodology is suitable for both scenarios of sufficient and "insufficient" follow-up, where the latter term refers to the situation where the right end point of the support of the censoring variable is strictly less than that of the time-to-event for the uncured observations. Extensive simulation studies and a real data analysis further illustrate the appropriateness of the methodology in its attempt to provide a supplementary tool for suitable analysis of randomized clinical trial results with cured subjects.
The event of interest in population-based cancer studies is usually death due to cancer. However, competing events that prevent the occurrence of the event of interest may be present. Relative survival is a commonly used measure that circumvents problems caused by inaccuracies in the cause of death information, incorporating other-cause mortality by matching cancer patients with individuals from the general population. Marginal estimates of relative survival summarise prognosis for a whole population, and contrasts of these (such as differences between subgroups) have a causal interpretation under certain assumptions. The causal inference literature in relative survival is scarce, with few application focussing on regression standardization that is the standard approach for obtaining marginal estimates in a modelling framework.

We propose two novel approaches to obtain marginal estimates of relative survival: i) inverse probability weighting (IPW) and ii) doubly robust standardization. In particular, we extend the IPW approach to use appropriate weights within the relative survival framework. With doubly robust standardization, a propensity score is first estimated as in the IPW approach. Then, the resulting weights are incorporated in a relative survival model alongside the exposure and all relevant covariates. With both methods, standard errors can be obtained by using either the delta method or M-estimation. The two methods outlined above as well as regression standardization are compared using a Monte Carlo simulation: we investigate the robustness of each approach to model misspecification, estimating both marginal estimates within exposed and unexposed as well as their difference.

Our results show that a higher degree of misspecification yields more biased estimates for regression standardization and IPW compared to doubly robust standardization, while the latter yields larger standard errors. All methods show less bias when the omitted variables are highly correlated with some of the variables included in the model. Finally, M-estimation tends to have better coverage compared to the delta method.

The methods we described can be extended to obtain several estimates of interest such as marginal all-cause survival and marginal crude probabilities, and can also be incorporated into mediation analysis approaches within the relative survival framework.

AUTHORS/INSTITUTIONS: E. Syriopoulou, M.J. Rutherford, P. Lambert, Health Sciences, University of Leicester, Leicester, UNITED KINGDOM|P. Lambert, Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, SWEDEN|
Drug sensitivity prediction with normal inverse Gaussian shrinkage informed by external data

Abstract Body: Precision cancer medicine relies heavily on accurate drug sensitivity prediction models. These models are often (partly) developed through simultaneous drug screening of tumour derived cell lines. The well-characterised molecular profiles of these cell lines provide ample opportunity for predictive modelling. These types of prediction problems entail modelling multivariate drug responses on high dimensional molecular feature sets in typically > 1000 cell lines. Typically the number of drugs are in the hundreds and the number of molecular features in the thousands. The dimensions of the problem require specialised models and estimation methods. In addition, external information on both the drugs and the features is often available, such as the target molecular pathways and developmental stages of the drugs.

We propose to model the drug responses through a Bayesian linear regression with shrinkage enforced through a normal inverse Gaussian prior. We let the prior depend on the external information on drugs and features, and estimate the model and its hyperparameters in an empirical-variational Bayes framework. Estimation of the hyperparameters leads to a data-dependent procedure that automatically weighs the informativeness of the external drug and feature information. The resulting Bayesian shrinkage model addresses the issues due to the high dimensionality of the problem.

To assess the performance of our model, we present a simulation study. The simulation study highlights the accurate hyperparameter estimation of the empirical Bayes procedure and shows an increase in predictive performance compared to models that do not use external information on the drugs and features. In addition, we demonstrate the usefulness of this model in the publicly available Genomics of Drug Sensitivity in Cancer data.

Authors/Institutions: M. Münch, Mathematical Institute, Leiden University, Leiden, Netherlands| M. Münch, M. van de Wiel, Epidemiology and Biostatistics, Amsterdam University Medical Centers, Amsterdam, Netherlands| M. van de Wiel, S. Richardson, G.G. Leday, MRC Biostatistics Unit, Cambridge University, Cambridge, United Kingdom|
We investigate calibration and assessment of prognostic rules when missing values are present in the predictors. This research has two key objectives. The first is to investigate how the calibration of any prediction rule can be combined with use of multiple imputation to account for missing predictor observations. The second objective is to propose such methods that can be implemented with current multiple imputation software in a pragmatic manner, while allowing for unbiased predictive assessment through validation on new observations for which outcome is not yet available.

We commence with a review of the methodological foundations of multiple imputation as a model estimation approach as opposed to a purely algorithmic description. We specifically contrast application of multiple imputation for parameter (effect) estimation with predictive calibration. Based on this review, two approaches are formulated, of which the second utilizes application of the classical Rubin’s rules for parameter estimation, while the first approach averages probabilities from models fitted on single imputations to directly approximate the predictive density for future observations. We present pragmatic implementations using current software which allow for validation and estimation of performance measures by cross-validation, as well as imputation of missing data in predictors on future data where outcome is missing by definition.

We will discuss application for both censored (survival) outcomes using the Cox model as well as binary outcome logistic regression prediction. Method performance is verified through application on two real datasets and simulation. Accuracy (Brier scores) and variance of predicted probabilities are investigated. Results show substantial reductions in variation of calibrated probabilities when using the first approach relative to use of Rubin’s rules. Furthermore, as compared to the prediction-averaging approach, variance levels from Rubin’s rules pooled models do not reduce to zero when numbers of imputations are increased. Irrespective of approach, numbers of imputations must be substantially increased from current practice for variation to reduce to acceptable levels for clinical application, with numbers between 100 and 1000 more realistic for reliable predictive calibration.

Authors/Institutions: B. Mertens, Biomedical Data Science, Leiden University Medical Centre, Leiden, Netherlands
Abstract Body:
Our dataset consists of around 4500 photoplethysmographic (PPG) signals acquired with smartphone cameras. Data pre-processing was performed by subtracting a moving average and dividing by the envelope (the absolute value of the Hilbert transform of the signal) in order to detrend and demodulate each signal. Then a peak detection algorithm was constructed, and poor signals with a low quality score representing incorrect demodulation and noise with low quality indexes were filtered out.

From the filtered data ca 100 features were extracted: the signal, the peak sequence, the difference between consecutive peaks sequence and the second derivative of the signal (SDPPG), etc. Our aim was to predict health ageing from these features. As outcome, we considered chronological age as well as dichotomized one (young vs old). For feature selection and to avoid overfitting, we repeatedly applied linear and logistic ridge regression 100 times over 100 different subsamples of the training dataset, and averaged the ranking of two regression results using absolute value of coefficients. This procedure resulted in selecting two features: Turning Point Ratio (basically a counter of local extrema) and 1 feature from the SDPPG group (the “a” wave - since it is the easiest to acquire and to interpret). These two features performed well for classification of young/old on the external test set (AUC=0.86), and by adding other covariates (smoking, gender, weight, etc) we achieved AUC=0.91.

Alternatively, we also considered a convolutional neural network (CNN) approach. Using chunks of 15 consecutive peaks of pre-processed signals as dataset we compared various types of CNN and topologies. The best results were achieved from a 12 layers ResNet (residual neural network) with comparable prediction performance of young/old classification, AUC=0.90.

Our results highlight that 2 properly selected features combined with the covariates can perform as well as using computer intensive CNNs, which may lead to a possibly faster analysis of wearables with few or no disadvantages.
Abstract Body: In the INFORM2 exploratory multinational phase I/II combination study of Nivolumab and Entinostat in children and adolescents with refractory high-risk malignancies, four biomarker-defined cohorts will be treated by the same treatment regime and response will be evaluated. The trial population is extremely small and sample size is governed by availability in the recruitment period. It is expected that up to 25 (in one arm, 29) patients will be recruited. A Bayesian adaptive design was chosen with an interim analysis to evaluate futility. Futility and efficacy will be assessed using Bayesian posterior probabilities, for trial design see [3].

In the final analysis, all four cohorts will be evaluated separately for efficacy. However, an exploratory analysis with borrowing between groups is planned. This analysis typically uses a Bayesian hierarchical model across the cohorts, as, e.g. promoted by Berry et al [1]. However, the potential benefit of borrowing information across cohorts has been questioned by, e.g. Freidlin and Korn [2] on basis of simulation studies. We have shown that no power gain is possible by borrowing from historical information when type I error should be controlled [4]. The same is true if information is borrowed across concurrent groups if no assumption can be made concerning the similarity of cohorts. However, if the cohorts are reasonably similar, power gains are possible. The circumstances under which this is true are investigated by analytical considerations.

References:

Authors/Institutions: A. Kopp-Schneider, S. Calderazzo, M. Wiesenfarth, Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, GERMANY
Abstract Body: Historical Functional Linear Models (HFLM) quantify associations between a functional predictor and functional outcome where the predictor is an exposure variable that occurs before, or at least concurrently with, the outcome. Prior work on the HFLM has largely focused on estimation of a surface that represents a time-varying association between the functional outcome and the functional exposure. This existing work has employed frequentist estimation methods, with little attention paid to formal inference or adjustment for multiple testing. In this work, we propose a new functional regression model that estimates the time-varying, lagged association between a functional outcome and a functional exposure. Building off of recently developed function-on-function regression methods, the model employs a novel wavelet packet decomposition of the exposure and outcome functions that allows the user to strictly enforce the temporal ordering of exposure and outcome, which is not possible with existing wavelet-based functional models. We take a fully Bayesian approach, which allows us to conduct formal inference on the time-varying lagged association, while adjusting for multiple testing. We investigate the operating characteristics of our wavelet-packet HFLM, as well as the resulting inference procedures, in simulation and use the model to analyze data on the impact of lagged exposure to particulate matter finer than 2.5μg on heart rate variability in a cohort of journeyman boilermakers over the course of a day’s shift.

Comparative effectiveness research relies heavily on the results of network meta-analysis (NMA) of randomized controlled trials (RCTs) to evaluate the efficacy and safety of multiple interventions. However, the NMA results may not generalize to all people in a target population of interest in which we want to make decisions regarding treatment implementation, because individual RCTs may not be representative of the target population. In this talk, we introduce NMA using Bayesian composite likelihood methods to estimate population treatment effects. First, to make RCT samples look like the population, we estimate the probability of being in RCTs given baseline characteristics and then calculate weights for all RCT participants. Second, these weights are integrated in a Bayesian network meta-analysis model with a composite likelihood function. These two steps can be conducted independently (two-step approach) or simultaneously (one-step approach), where the latter fully incorporates the uncertainty of weights. We show simulation study results to validate and compare the two proposed approaches with a conventional NMA model. We apply these methods to generalize NMA comparing 5 different antipsychotics treatments and placebo on schizophrenia to the US population of adults with schizophrenia who present to usual care settings.
Abstract Body: Introduction: Quantitative Immunofluorescence (QIF) is used for immunohistochemistry (IHC) quantification of proteins that serve as cancer biomarkers. A common application of QIF-IHC is analysis of protein expressions in tissue microarrays (TMAs) that incorporate tumor tissues from large cohorts of patients. This technology combined with integrated image analysis allows for high-throughput measurement of proteins in individual cancer cells. However, only the mean signal intensity (MSI) of the protein expression across cancer cells is usually considered for developing protein biomarkers.

Methods: We propose a new approach for developing biomarkers using the information on spatial distribution of cellular signal intensity (CSI) of protein expression in cancer cell population. We view the protein QIF expression levels as marks in marked point process of cancer cells in the tissue and develop new spatial index predictors of clinical outcomes based on conditional mean and conditional variance of the marked point process. These characteristics of the marked point process are estimated using nonparametric kernel density estimates of under the assumption of second-order intensity reweighted stationary processes, which allow accommodating marked point processes with variably inhomogeneous point patterns of cancer cells.

Results: A simulation study demonstrates the ability of the proposed spatial indices to discriminate marked point patterns similar to the ones observed for protein expression in cancer cells. The utility of new spatial index protein biomarkers is investigated and compared to the standard MSI predictors using the protein expressions in tissue microarrays (TMAs) incorporating tumor tissues from over 1,000 breast cancer patients.

Discussion: The new approach provides new insight into standard IHC protein biomarkers and identifies novel biomarkers that do not have a prognostic value if only the mean signal intensity is considered.

Keywords: Nonparametric statistics, Personalized medicine, Risk Prediction; Statistical machine learning; Other-Marked point process, Marked point patterns, Second-order characteristics of marked point processes Microscopic image analysis, Quantitative Pathology

AUTHORS/INSTITUTIONS: I. Chervoneva, M. Yi, Thomas Jefferson University, Philadelphia, Pennsylvania, UNITED STATES|H. Rui, Medical College of Wisconsin Cancer Center, Milwaukee, Wisconsin, UNITED STATES
Abstract Body: In biomedical studies, markers are very useful to monitor the evolution of a disease over time. Studying their trajectory gives insight on the disease natural history. In dementia studies, it is interesting to assess the evolution of cognitive abilities during the pre-diagnosis phase that can last around fifteen years. If the cognitive decline trajectories are quite heterogeneous, it is known that, for cases, trajectories are nonlinear while normal cognitive ageing is generally smoother. Assessing the difference between pathologic and normal cognitive ageing is useful for understanding the pathological process leading to dementia. In this work, we are mainly interested in identifying and comparing the time of differentiation between normal and pathological decline for various cognitive functions.

We propose a semi latent class random changepoint model that models cognitive decline trajectories for cases and controls from a nested case control study built from an observational cohort. In this design, each case is matched with one or several controls free of dementia at the diagnosis visit of the case. The model has two classes: a first class where the trajectory follows a linear mixed model and a second class where the trajectory follows the same linear trajectory as the first class up to a certain time, the time of differentiation, after which an accelerated nonlinear decline happens. All cases are assumed to belong to the second class while controls may belong to both classes. Indeed, as a control may develop dementia after the diagnosis visit of the matched case, it is not realistic to assume that all controls are on the linear phase. Parameters are estimated by maximum likelihood using a Levenberg-Marquardt algorithm and Gaussian quadrature for computing the integral on the random changepoint.

The model is estimated on data from the French cohort PAQUID to compare the time of differentiation of several cognitive tests.

Authors/Institutions: C. Segalas, H. Jacqmin-Gadda, Bordeaux Population Health, INSERM U1219, Bordeaux, FRANCE
The global increase in antimicrobial resistance (AMR) represents a major public health crisis. The current project concerns analysis of collateral effects of AMR, which occur when resistance against one antibiotic leads to either increased resistance, or increased sensitivity against a second antibiotic. Knowledge of such collateral effects is of clinical relevance as these may enable the design of antibiotic dosing schedules that limit AMR. Sensitivity or resistance of pathogens against antibiotics is quantified using the minimum inhibitory concentration (MIC). The MIC reflects the lowest concentration that leads to inhibition of growth, and is typically measured at discrete and exponentially increasing antibiotic concentrations, with an underlying continuous distribution. MIC data for most antibiotics in hospital-isolated pathogens is collected in population surveillance studies. Statistical methodologies to effectively identify both the occurrence, effect size and effect directionality of collateral effects of AMR in pathogens that are associated with clinical infections are however lacking. We describe a novel goodness-of-fit measure to address this challenge.

We consider the situation where the effect of changes in MIC for two antibiotics is tested as a goodness-of-fit measure. All available MIC observations for an antibiotic (A) are used to estimate a marginal MIC distribution. A subset of these MIC observations, with a specified threshold value for the MIC of an antibiotic B indicating resistance, is used to estimate the conditional distribution of antibiotic A. The difference between the means is tested using a dependent sample T-test, where all observations in the conditional sample are contained in the marginal sample. Antibiotic B can also be tested conditionally on resistance to antibiotic A, which can yield very different results, setting this test apart from simple correlation tests. We apply this measure to a large clinical AMR surveillance database containing MIC data measured in the pathogen Escherichia coli. We identify several antibiotic combinations that show distinct collateral effects. We conclude that the developed methodology is of relevance for the systematic identification of collateral effects in other pathogens and population surveillance databases.
Abstract Body: Quality patient care requires comprehensive health care data from a broad set of sources. Electronic medical records (EMR) are increasingly distributed across many sources as our nation moves into an era of electronic health record systems. But EMR data are often from independent databases without a common patient identifier, the lack of which impedes data aggregation, causes waste (e.g., tests repeated unnecessarily), affects patient care and hinders research. Record Linkage is the first requisite step before effective and efficient patient care and research. Absent a unique universal patient identifier, linkage of patient records is a non-trial task. In addition, the ubiquity of missing data in EMR poses further challenges in record linkage. We will present our novel solution to patient record linkage in the presence of missing data in a real-world big data setting. The Indianapolis Network for Patient Care (INPC) was created in 1995 with the goal of providing clinical information at the point of patient care. It houses clinical data from over 80 hospitals, public health departments, local laboratories and imaging centers, surgical centers, and a few large-group practices closely tied to hospital systems, for approximately 13.4 million unique patients. Our novel approach was applied to four linking tasks representative of real-work clinical and public health scenarios. Performance measures such as sensitivity, positive predictive value and the F-score were computed for the novel approach and the default method for missing data in record linkage. Our novel method conferred an absolute increase in the overall F-score from 2% to 24% with an absolute increase in the sensitivity from 5% to 40%.

AUTHORS/INSTITUTIONS: X. Li, Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, UNITED STATES|S. Grannis, Regenstrief Institute, Indianapolis, Indiana, UNITED STATES|H. Xu, Biostatistics, Indiana University School of Public Health, Indianapolis, Indiana, UNITED STATES|
The field of movement ecology has experienced unprecedented growth in the last decade, due to technological advances that have brought large volumes of tracking data from various devices (e.g. GPS, accelerometers, video cameras), and analytical and programming tools that aid data processing and analysis. We identified >8000 scientific publications from the last decade, and used a text mining procedure to review the research done in movement ecology in several dimensions (research topics, species studied and methodological tools). A “dictionary” approach was used for aspects like software or biologging device used, a list of possible software/devices was known a priori. Dictionaries are list of terms (e.g. “R Software”, “R Development Core”), grouped at a higher level under a set of categories (e.g. the category “R”). When a term in the dictionary was found in the text of a manuscript (e.g. in Material and Methods or the abstract), the manuscript was associated to the category the term corresponded to. Research topics, however, were not defined a priori. To identify the unknown and hidden topics in the publications, we fitted Latent Dirichlet Allocation models to the abstracts, which are three-level (topic, words, abstract) hierarchical Bayesian models for documents. Topics were defined by the model as a mixture of words, and as each abstract could be composed of one or more topics, the choice of words in the abstract would reflect the topics behind it; the probability of a given word appearing in an abstract would depend on the topic the abstract is addressing. In this presentation, we describe the methodology used to analyze the state of the field of movement ecology and present results based on the dictionary and topic modeling approaches.

AUTHORS/INSTITUTIONS: R. Joo, M.E. Boone, S. Picardi, M. Basille, Department of Wildlife Ecology, University of Florida, Davie, Florida, UNITED STATES|T.A. Clay, S.C. Patrick, School of Environmental Sciences, University of Liverpool, Liverpool, UNITED KINGDOM|V. Romero, Universidad de Lima, Lima, PERU
Several model comparison techniques exist to select a best model from a set of candidate models. This study explores the performance of model comparison statistics among several Bayesian software packages that are often used for spatially discrete disease modelling: the deviance information criterion (DIC), the Watanabe-Akaike information criterion (WAIC) and the log marginal predictive likelihood (LMPL). We focus on the software packages CARBayes, OpenBUGS, NIMBLE and Stan, in which we fit Poisson models to disease incidence outcomes with intrinsic conditional autoregressive, convolution conditional autoregressive and log-normal error terms. From three data analyses, that differ in the number of areal units and disease prevalence, we learn important disparities in model selection. Based on these conclusions, we provide recommendations on the optimal use of model comparison statistics for all kind of applications.

**Authors/Institutions:** M. Vranckx, T. Neyens, C. Faes, Hasselt University, Diepenbeek, BELGIUM | M. Vranckx, T. Neyens, C. Faes, Data Science Institute (DSI), Diepenbeek, BELGIUM | M. Vranckx, T. Neyens, C. Faes, Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Diepenbeek, BELGIUM
Abstract Body: In air pollution epidemiology, measurements of relevant exposure concentrations are typically made at point locations, resulting in spatial misalignment between the exposure data and health outcomes aggregated at the area level. To obtain values that match the spatial units of the health data, observations can be averaged directly or prediction models developed. We present a framework for evaluating the error in aggregating exposure concentrations to the area unit. We present estimators of mean squared error that can be used for model selection. We find that exposure prediction models, even when misspecified, outperform monitor averages in settings with realistic numbers of monitors and that important reductions in error of the health effect estimate can be obtained when restricting to areas with a monitor. In an analysis of long-term particulate matter concentrations across the United States, we estimated the error of the prediction model approach to be less than that of the monitor averaging approach on average across counties. We present health effect estimates about particulate matter exposure and pediatric asthma morbidity in the Medicaid population using each approach. Our findings support the use of a prediction model for estimating area-wide averages, even when restricting to areas that contain a monitor.
Abstract: We propose an extension of the COM-Poisson model to jointly model the mean and the dispersion as functions of covariates taking into account, possibly, under- and overdispersion in the same count data set. Estimation and inference are based on the likelihood paradigm. Results from a simulation study show that the maximum likelihood estimators are consistent and unbiased for both mean and dispersion parameters. The methodology is illustrated with the analysis of a data set. The R codes and data set are available online.
CONTROL ID: 3367815
TITLE: A new sequential monitoring method for event rate in a clinical trial

ABSTRACT BODY:

Abstract Body: In this talk, we introduce a new continuous monitoring method for the event rate of time-to-event data when patients enter the clinical trial in a staggered fashion. Built on a sequential probability ratio test using boundaries derived from the nonlinear renewal theory with stationary perturbations (Kim and Woodroofe 2003), the sequential method uses both the counts of primary events and cumulative time of patients on trial. The monitoring gives an early warning if the target event rate is unlikely to be achieved by the end of the follow-up period. If necessary, this method can be used to suggest an extension of follow-up period to achieve the target rate with specified probability. We illustrate the method using the data from a Phase III clinical trial.

AUTHORS/INSTITUTIONS: D. Kim, National Institutes of Health, Bethesda, Maryland, UNITED STATES]
Environmental exposures can regulate intermediate molecular phenotypes, such as gene expression, by different mechanisms and thereby lead to various health outcomes. It is of significant scientific interest to unravel the role of potentially high-dimensional intermediate phenotypes in the relationship between environmental exposure and traits. Mediation analysis is an important tool for investigating such relationships. However, it has mainly focused on low-dimensional settings, and there is a lack of a good measure of the total mediation effect. Here, we extend an R-squared (Rsq) effect size measure, originally proposed in the single-mediator setting, to the moderate- and high-dimensional mediator settings in the mixed model framework. Based on extensive simulations, we compare our measure and estimation procedure with several frequently used mediation measures, including product, proportion, and ratio measures. Our Rsq measure has small bias and variance under the correctly specified model. To mitigate potential bias induced by non-mediators, we examine two variable selection procedures, i.e., iterative sure independence screening and false discovery rate control, to exclude the non-mediators. We establish the consistency of the proposed estimation procedures and introduce a resampling-based confidence interval. By applying the proposed estimation procedure, we find that more than half of the aging-related variations in systolic blood pressure can be explained by gene expression profiles in the Framingham Heart Study of 1,711 individuals. We have implemented the proposed method in an R package "RsqMed".
SNP-based heritability of a trait is measured by the proportion of variance explained by additive effects of the observed single nucleotide polymorphisms (SNPs) over the entire genome. The Genome-wide complex trait analysis (GCTA) approach is now being routinely used to estimate SNP-based heritability for many complex traits. The basic concept behind this approach is to fit the effects of all the SNPs as random effects in a mixed linear model, where the variance of these genetic effects attributes to the heritability of the trait. The approach models the genotypic covariances among individuals in the sample: these covariances are captured by estimating a genetic relationship matrix (GRM). Heritability is estimated by the restricted maximum likelihood (REML) approach and the estimation relies heavily on the GRM estimated from the SNPs. Presence of subtle population substructures in the data could significantly impact the GRM estimation and may introduce bias in the heritability estimation. The common practice of accounting for such population substructure is to adjust for the top few principal components of the GRM as fixed effects in the mixed linear model. However, such adjustments often fail to correct for the bias due to population substructure. Here we propose an alternative way of estimating heritability in multi-ethnic studies through an estimating equation approach. Our proposed approach introduces adjustments for the population stratification in the estimating equation and allows for the additive genetic variance and residual variance to vary by ethnicity. We study the performance of our approach and the GCTA-REML approach with or without PC adjustments in presence of population stratification through extensive simulation studies. We estimate the heritability of height, BMI and other anthropometric traits in the UK Biobank cohort to investigate the impact of subtle population substructure on SNP-based heritability estimation for moderately and highly heritable traits.

**AUTHORS/INSTITUTIONS:** S. Basu, S. Seal, Z. Lin, Biostatistics, University of Minnesota, Minneapolis, Minnesota, UNITED STATES; A. Min, Statistics, University of Washington, Seattle, Washington, UNITED STATES
Understanding the interplay between different types of cells and their immediate environment is critical for understanding the mechanisms of cells themselves and their function in the context of human diseases. Recent advances in high-parameter imaging cytometry technologies have fundamentally revolutionized our ability to observe these complex cellular relationships providing an unprecedented characterisation of cellular heterogeneity in a tissue environment.

Imaging cytometry data facilitates the identification of spatial organization of cell-types. That is, imaging cytometry data can be used to identify evidence that cell-types are aggregating or colocalising with each other. Established immunohistochemistry protocols only enable the visualization of a few cell-types concurrently and as such, analytical methods have been developed to identify the presence of pairwise spatial relationships between cell-types. However, high-parameter imaging cytometry technologies have now made it possible to simultaneously quantify hundreds or thousands of pairwise interactions between cell subsets. This has produced an urgent and exciting need to develop analytical frameworks that are necessary to tease apart and prioritise these high-dimensional interactions for further investigation.

In this presentation, I will introduce a novel methodology for identifying consistent spatial organisation of multiple cell-types in an unsupervised way. In short, our method clusters local indicators of spatial association (LISA) functions to enable the characterization of interactions between cell-types in contrast to traditional pairwise analysis. LISA curves are a localised summary of an L-function derived from a Poisson point process model and have been used in varied contexts to identify landmines or denoise images for text recognition. I will demonstrate the effectiveness of this framework on data generated from multiple high-parameter imaging cytometry assays including CODEX, CycIF and IMC. These illustrations showcase the utility of these new technologies and demonstrate that analysing cell-type interactions as a complex system, as opposed to a battery of pairwise tests, provides clear interpretational advantages.

AUTHORS/INSTITUTIONS: E. Patrick, School of Mathematics and Statistics, The University of Sydney, The University of Sydney, New South Wales, AUSTRALIA|E. Patrick, Westmead Institute for Medical Research, Westmead, New South Wales, AUSTRALIA|
In recent years there has been a growing interest in designs that incorporate data collected outside of experimental studies, also known as real world data, or data from completed trials. In this talk I will discuss two uses of external data (previous trials and real world data) in the design and analyses of clinical studies. Firstly, I will introduce a Bayesian hybrid platform design that leverages external data via a non-parametric Bayesian model averaging approach. I will show the usefulness of this hybrid design using a collection of phase II and III trials of patients with glioblastoma for an immunotherapy agent. Secondly, I will discuss validation techniques for the accuracy of clinical outcome predictions obtained leveraging external data, and compare their performance with other predictive values that do not incorporate external data. I will illustrate the latter strategy with a library of patient-level data from glioblastoma trials.
Abstract Body: With the advent of precision medicine there has been an increasing attention on designs which accelerate drug development relying on additional sources of information. We incorporate auxiliary endpoints measured during a trial in a novel Bayesian theoretic decision framework which allows for efficient allocation of resources. Such endpoints are currently used as surrogate for primary outcomes if considered valid or discarded from the analysis, even when still potentially informative of the treatment effect. Our approach combine the use of interpretable utility functions with frequentist criteria. We select the optimal design which satisfies a set of operating characteristic defined on the primary outcome, being therefore robust to the validity of the auxiliary endpoints included in the design. We consider the problem of selecting subsets of multiple auxiliary outcomes, assessing their importance via suitable metrics. Some theoretical properties are described, and we illustrate the behavior of the proposed framework via group sequential designs.

AUTHORS/INSTITUTIONS: M. Russo, Harvard-MIT Center for Regulatory Science (CRS), Harvard University, Boston, Massachusetts, UNITED STATES|L. Trippa, S. Ventz, Biostatistics, Harvard University, Boston, Massachusetts, UNITED STATES|M. Russo, L. Trippa, S. Ventz, Data Sciences, Dana Farber Cancer Institute, Boston, Massachusetts, UNITED STATES
Abstract Body: In this study we introduce a likelihood-based method, via the Weibull and piecewise exponential distributions, capable of accommodating the dependence between failure and censoring times. The methodology is developed for the analysis of clustered survival data and it assumes that failure and censoring times are mutually independent conditional on a latent frailty. The dependent censoring mechanism is accounted through the frailty effect and this is accomplished by means of a key parameter accommodating the correlation between failure and censored observations. The full specification of the likelihood in our work simplifies the inference procedures with respect to Huang and Wolfe since it reduces the computation burden of working with the profile likelihood. In addition, the assumptions made for the baseline distributions lead to models with continuous survival functions. In order to carry out inferences, we devise a Monte Carlo EM algorithm. The performance of the proposed models is investigated through a simulation study. Finally, we explore a real application involving patients from the Dialysis Outcomes and Practice Patterns Study observed between 1996 and 2015.
Abstract Body: DNA testing became compulsory in New South Wales, Australia in 2001. To date, there is limited evidence in terms of population estimates to show that DNA testing has had a positive impact on reducing certain types of crime. This study used techniques developed in wildlife biology to estimate the population of offenders who engaged in burglary in New South Wales during two time periods 1999/2000 and 2003/2004 respectively. Data comprised records from the New South Wales Bureau of Crime Statistics and Research re-offending database. Capture-recapture modelling estimated the population of burglary offenders to have been approximately 44,000 in 1999/2000 and 25,500 in 2003/2004. There was evidence of two distinct groups of offenders with high and low rate of offending respectively. In general, high rate offenders are caught and convicted but low rate offenders escape conviction. The decrease of 42% in the population size of burglars from 1999/2000 to 2003/2004 indicated that the introduction of DNA testing may have contributed to a decrease in burglary related activities.

Authors/Institutions: J. Thandrayen, Research School of Population Health, Australian National University, Acton, Australian Capital Territory, AUSTRALIA|I. Hudson, School of Mathematical Sciences, RMIT University, Melbourne, Victoria, AUSTRALIA|D. Weatherburn, New South Wales Bureau of Crime Statistics and Research, Sydney, New South Wales, AUSTRALIA|
Abstract Body: The reparameterized t-Student distribution considered in this paper has the same covariance matrix as the normal distribution, enabling a direct comparison between them. Another advantage is that the values for the parameter of form is bounded, and under the assumption that the second moment exists, it allows the feasibility of the parameters estimation, by maximizing the concentrated log-likelihood and the implementation of the EM algorithm. The main contribution is to obtain an appropriate perturbation scheme to the covariance matrix, and developing diagnostic gradient statistics to identify the presence of influential observations that can interfere in the parameter estimation, model selection and prediction. Some identifiability conditions are presented. A soybean yield real data is analysed. The predicted values were obtained using kriging. It shows that the reparameterized t-Student distribution is an alternative to analyse data with outliers, because it is less sensitive to the presence of this type of observations.
Abstract Body: The mRNA alternative splicing events enable a single gene to produce different proteins in eukaryotes, and they have been shown to affect various gene functions, and eventually disease. The alternative splicing can not only add or skip entire exons, but can also vary exon boundaries by selecting different nucleotides as splice sites. Oxford Nanopore sequencing produces long reads that have natural advantages for characterising alternative splicing events. Nanopore sequencing records changes in electrical current when a DNA or RNA strand is traversing through a pore. This raw signal, known as a squiggle, is then basecalled by computational methods. However, due to the high error rate in the basecalling process, it is challenging to accurately identify splice sites using the basecalled reads. One solution is polishing splice sites identified by the Nanopore reads by using short reads from the same samples, but matched short reads add costs and are often not available. Therefore, a method that could accurately identify splice sites solely from Nanopore sequencing data would have numerous advantages. In this talk, we will present new methods that use the raw signals (the squiggles), in addition to the basecalled reads, to characterise splice sites. We tested our methods using synthetic mRNAs with known splice sites, demonstrating our methods can improve on methods that use only the basecalled reads. We will also illustrate how outputs from the proposed methods can be used in downstream alternative splicing analysis such as differential alternative splicing analysis.

AUTHORS/INSTITUTIONS: Y. You, H. Shim, School of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, AUSTRALIA|Y. You, H. Shim, Melbourne Integrative Genomics, University of Melbourne, Melbourne, Victoria, AUSTRALIA|M. Clark, Anatomy and Neuroscience, School of Biomedical Sciences, University of Melbourne, Melbourne, Victoria, AUSTRALIA|
Bootstrap of Association Matrices (BAM): A Robust Method for Integrated Analysis of Multiple Omic Data Sets with Multiple Clinical Outcomes

Abstract Body: Modern biomedical research studies routinely collect multiple matrices of data on thousands that each contain thousands to millions of variables in order to explore the associations of these data matrices with multiple pharmacologic and/or clinical outcomes. For example, in oncology research, there may be a data matrix for each of several types of molecular data such as DNA genotype, DNA abnormalities, or RNA expression levels. The objective is to identify genes for which one or more of these molecular data forms are robustly associated with one or more clinical characteristics or outcomes such as risk of disease development, disease histology, response to chemotherapy, time to relapse, and time to death.

Here, bootstrap of association matrices (BAM) is introduced as a robust method that can be applied to a broad class of these types of research studies. First, BAM computes estimates of a matrix of association parameters for each clinical outcome with each molecular variable for the observed data set. Next, BAM computes a series of bootstrap estimates of this association matrix for each of many bootstrap data sets obtained by resampling subjects with replacement. This bootstrap procedure produces a cloud of points in multivariate space; each point representing the association vector for one bootstrap. Finally, for each gene with a predefined association submatrix of scientific interest, BAM uses a recursive peeling algorithm to quantify statistical significance by characterizing the position of the origin (null) relative to the point cloud of bootstrap association estimates.

BAM possesses several practical advantages over other widely used methods for integration of multi-omic data with multi-endpoint data. BAM provides a more elegant approach to adjust for clinically relevant covariates than do widely used permutation methods. BAM does not require specification and evaluation of priors like Bayesian methods or specification of data integration weights like projection onto the most interesting statistical evidence.

The performance of BAM will be evaluated and compared to that of other methods in simulation studies and in the analysis of example data sets from pediatric cancer research. BAM will be made freely available as an R package on GitHub and/or CRAN.

Authors/Institutions: S.B. Pounds, L. Shi, Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, UNITED STATES|X. Cao, Preventative Medicine, University of Tennessee Health Science Center, Memphis, Tennessee, UNITED STATES|
Subject to competing risks. A few researchers have observed that associations among multivariate failure times subject to competing risks can be decomposed into two elements: the association between the probabilities that the individuals experience a specific failure type, and the association between times to first failure regardless of cause. Bandeen-Roche and Liang (2002) demonstrated that under certain assumptions, the CCSHR can be expressed as a multiplication of the ordinary CHR for association between times to first failure, and a factor representing the association between failure causes. They also utilized the ordinary CHR for association between times to first failure. The previous approach to estimating the decompositions described above either have been parametric (Bandeen-Roche & Liang, 2002) or modeled time dependence as piecewise constant.

The decompositions described above distinguish associations between causes of failure from associations between overall propensity to fail. In this paper, we aimed to augment the available methodology to achieve such insights by enabling completely nonparametric estimation of each component in the CCSHR decomposition. To this end, we developed methods to estimate both failure-time and failure-cause components of the CCSHR by smoothing.
Accounting for Calibration Drift due to Improvements in Baseline Survival during Prognostic Model Development

Sarah Booth¹, Paul C Lambert¹,², Mark J Rutherford¹
¹Department of Health Sciences, University of Leicester, UK
²Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden

Abstract

Introduction

Prognostic models which are used to produce long-term survival predictions are often developed using datasets which cover a long diagnosis period. If survival is continually improving during this time, it can lead to predictions which under-estimate the survival of recently diagnosed patients. The developments here focus on models which produce survival predictions following a cancer diagnosis, however these methods can be generalised to other outcomes.

Methods

An extensive comparison of approaches for accounting for improvements in survival was performed. This included modelling calendar time directly by including the year of diagnosis as a predictor, and if required, allowing for non-proportional hazards and interaction terms. Alternative period analysis based methods involved using delayed-entry techniques to analyse the most recent subset of data and temporal recalibration which uses this subset to re-estimate the baseline hazard of standard prognostic models. We compare approaches in a simulation setting under a range of plausible assumptions and further show examples of models for colon cancer developed using US population-based registry data from the SEER database.

Results

Accounting for changes in baseline survival often improved the calibration of the predictions for new patients compared to the standard model. However, when time-dependent effects for the year of diagnosis were included, this sometimes resulted in inaccurate survival predictions. This is likely due to the most recently diagnosed patients having a limited amount of follow-up which leads to the inappropriate extrapolation of a non-linear interaction effect based on the long-term information from patients who were diagnosed earlier.

Conclusion

Care should be taken when modelling complex trends in calendar time; particularly when allowing the effect of calendar time to be time-dependent. A more conservative approach to produce up-to-date survival predictions can be achieved using period analysis based methods as less extrapolation is required.

Authors/Institutions: S. Booth, P.C. Lambert, M.J. Rutherford, Department of Health Sciences, University of Leicester, Leicester, UNITED KINGDOM|P.C. Lambert, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, SWEDEN|
Electronic health records (EHRs) offer great promises of advancing precision medicine and, at the same time, present significant analytical challenges. Particularly, it is often the case that patient-level data in EHRs cannot be shared across institutions (data sources) due to government regulations and/or institutional policies. As a result, there is a growing interest in federated learning from multiple EHRs databases that does not require sharing patient-level data. To tackle such challenges, we propose a new communication efficient algorithm that aggregates the local information about parameters in the (generalized) linear model. Our approach requires only a single one-way communication from all sites to the central site, where the analysis is taking place, and is capable of utilizing the local data in the central site to improve estimation and inference. We investigate the operating characteristics of the proposed method and evaluate its performance by simulations in comparison with several recently developed methods.

C. Chang, Z. Bu, Q. Long, University of Pennsylvania, Iowa City, Iowa, UNITED STATES
Application of multiple testing procedures for identifying multi-morbidities preceding a condition of interest in big health-administrative data

Background: Multiple testing procedures (MTP) are gaining increasing popularity in various fields of biostatistics, especially in statistical genetics. However, in epidemiologic literature utilizing health-administrative data, there are few studies that applied MTP and discussed their applications and challenges. The objective of our study is to introduce and apply the statistical methods of MTP in association studies and predictive modelling in the context of multi-morbidities using health-administrative data by using MTP to find a subset of variables that have a high likelihood of being associated with the outcome of interest.

Methods: 2600 International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10) codes were used to assess the associations between the condition of interest and multi-morbidities on a matched dataset. For an illustration of the statistical tools outlined in this paper, we used a dataset of more than two hundred thousand patients with traumatic brain injury (TBI). McNemar tests were conducted on each of the 2600 ICD-10 codes and appropriate multiple testing adjustments were applied using the Benjamini-Yekutieli procedure. To study the direction of association between the ICD-10 codes and TBI event, odds ratios with 95% confidence intervals were constructed.

Results: The Benjamini-Yekutieli procedure captured 684 ICD-10 codes, out of the 2600, as codes relevant to a TBI event.

Conclusions: Our results illustrate the use of innovative statistical tools on big health-administrative datasets for data mining and dimension reduction.

Authors/Institutions: S. Jana, Indian Institute of Management, Nagpur, Nagpur, Maharashtra, INDIA|M. Sutton, A. Colantonio, M. Escobar, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, CANADA|T. Mollayeva, V. Chan, A. Colantonio, Toronto Rehabilitation Institute, University Health Network, Toronto, Ontario, CANADA|A. Colantonio, Rehabilitation Sciences Institute, Faculty of Medicine, University of Toronto, Toronto, Ontario, CANADA
Background. Designing trials to reduce treatment duration is important in several therapeutic areas, including TB and antibiotics. We recently proposed a new randomized trial design to overcome some of the limitations of standard two-arm non-inferiority trials. This DURATIONS design involves randomising patients to a moderate number of duration arms, and modelling the so called 'duration-response curve'. In this talk I will explore the operating characteristics of different statistical methods of drawing inference from the estimated curve.

Methods. Our first estimation target is the shortest duration non-inferior to the control one within a specific risk difference margin. We compare different methods of estimating this quantity, including using model confidence bands, the delta method and bootstrap. We then explore the generalisability of results to estimation targets which focus on cure rate, risk ratio and gradient of the curve.

Results. We show through simulations that, in most scenarios and for most of the estimation targets, using the bootstrap to estimate variability around the target duration leads to good results for DURATIONS design-appropriate quantities analogous to power and type 1 error. Using model confidence bands is not recommended, while the delta method leads to inflated type 1 error in some scenarios, particularly when the optimal duration is very close to one of the randomized durations.

Conclusions. Using the bootstrap to estimate the optimal duration in a DURATIONS design has good operating characteristics in a wide range of scenarios, and can be used with confidence by researchers wishing to design a DURATIONS trial to reduce treatment duration. Uncertainty around several different targets can be estimated with this bootstrap approach.

AUTHORS/INSTITUTIONS: M. Quartagno, J.R. Carpenter, M. Parmar, MRC-CTU at UCL, London, UNITED KINGDOM; J.R. Carpenter, Medical Statistics, London School of Hygiene & Tropical Medicine, London, UNITED KINGDOM;
Predicting clinical outcomes using multi-omics data, which consists for example of genomics, proteomics or metabolomics data, requires statistical methods that can deal with the situation of a huge number of covariates versus a relatively small number of observations. One option is the application of penalized regression like the Lasso, which results in sparse models because variable selection is part of the model building process.

In contrast to the standard Lasso model, the IPF-Lasso individually penalizes the various groups (modalities) of multi-omics data based on a set of manually entered candidate penalty-factors. However, this advantage has to be paid for with exponentially increasing computation time when there are more than two or three modalities involved.

In order to avoid this problem, we propose an extension of the IPF-Lasso called adaptive IPF-Lasso which follows a two-step approach:

In step 1 the penalty factors are generated based on the outcome of a penalized regression model, non-relevant modalities are removed from the input data, and the penalty factors are fed into the original IPF-Lasso in step 2. Several variants of the adaptive IPF-Lasso model are evaluated and compared to the original IPF-Lasso and other competitors by analyzing simulated and real data sets. The results show that for a given data set, the new adaptive IPF-Lasso needs only a fraction of the computation time of the original IPF-Lasso model and nevertheless can compete in terms of prediction performance.

An R-function for implementing the adaptive IPF-Lasso is included into version 1.1 of the R-package ipflasso.

**AUTHORS/INSTITUTIONS:** G. Schulze, Independent Researcher, Buehl, GERMANY | A. Boulesteix, R. Hornung, Department of Medical Information Processing, Biometry and Epidemiology (IBE), Ludwig-Maximilians-University Munich, Munich, GERMANY |
Abstract: In this work we provide an in-depth analysis of the discursive debate that has developed in Italy around compulsory vaccination of children in school age and that culminated, amidst a fierce controversy, in 2017 with the emanation of a decree law (the so-called Lorenzin law) by the Italian government. The decree not only increased the number of compulsory vaccinations from 6 to 12, but also imposed a series of disciplinary actions targeted at parents that did not vaccinate their children.

The study is a longitudinal study of the discursive debate around compulsory vaccination in the Italian context. The chosen timeframe is of roughly two years before and after the emanation of the Lorenzin decree (June 2017), i.e., from 1 January 2015 to 21st January 2019. To reconstruct the discursive debate around compulsory vaccination, we have opted for ANSA as source of data, the most important news agency in Italy providing news to the great majority of Italian newspapers and TV news programs. The reasons for such choice are several. First, we aim at reconstructing the debate around this issue by including the voices of all those actors that have contributed to it in the public space. Second, ANSA reports statements verbatim without any interpretative filter and, therefore, allows to avoid some of the typical limitations of media studies in which it is impossible to distinguish what is said by journalists and what by a specific actor. Third, day by day, ANSA, unlike social media, reports statements of a certain length or, at least, without any limitation in length. As we envisage to conduct argumentation analysis on these texts we believe ANSA statements are more conducive than brief twitters to elaborate arguments.

We created a corpus of 16324 documents, extracted from the ANSA database and annotated and manipulated with Python so to analyse subcorpus on the basis of the author, organization or date and as such to compare the statements produced by different subsets of actors or in different time periods.

We explore the effectiveness of a topic modelling analysis of the corpus with Latent Dirichlet Allocation. We aim on one side at identifying the main topics of the discursive debate and on the other side at using such topics to identify possible coalitions of actors, sharing a similar verbal approach in the debate.

Authors/Institutions: R. Graziani, A. Compagni, R. Florio, L. Bocconi University, Milan, ITALY
Simultaneous modeling of data arising from multiple ordered layers provides insight into the holistic picture of the interactive system and the flow of information. Such data structures are now routinely prevalent and include multi-platform `omics data from different molecular levels such as genomic, transcriptomic, and proteomic platforms that are measured on the same set of samples. We developed a general Bayesian framework to construct high-dimensional graphical models that allow both directed and undirected edges within and between variables at different hierarchical levels, and causal mediation analysis framework for the multi-layered structure with continuous, binary and survival responses.

M. Ha, J. Long, Biostatistics, MD Anderson Cancer Center, Houston, Texas, UNITED STATES|F. Stingo, Statistics and Computer science, University of Florence, Florence, ITALY|V. Baladandayuthapni, Biostatistics, University of Michigan, Ann Arbor, Michigan, UNITED STATES
Abstract Body: In this report we develop two Bayesian methodologies for the adjustment of two types of Hidden Markov Model (HMM in short) for time series of counting data. The premier model is the Poisson HMM (PHMM) and the second one is a Zero-Inflated Poisson-HMM (ZIP-HMM). The estimation in those models is done using Monte Carlo Hamiltonian and No-U-Turn sampler methods, and, the Bridge sampler is used to solve the unresolved problem of selecting the best model from the Bayesian approach. Finally, we present two applications to real data, in the first one the PHMM is used for modelling annual number of homicides in Colombia from 1960 to 2018, and next we use a ZIP-HMM for modelling the monthly time series of large forest fires in Colombia in the period January 2002 to December 2016.

From the results, we conclude that the proposed methodology for PHMM get a better adjustment than classical approach. The performance of the proposed ZIP-HMM is similar to the classical approach.
Abstract Body: For analytical similarity assessment of a given critical quality attribute between a proposed biosimilar (test) product and an innovative (reference) biological product, FDA recommended an equivalence test with a margin of 1.5 $\sigma_R$ (standard deviation of the reference product) be performed. The FDA recommended similarity margin has been criticized by many authors in the literature due to its inflexibility. In this article, we proposed an equivalence test with flexible margin for controlling possible inflation/deflation of the variability associated with the response. The performance of the proposed equivalence test with flexible margin is evaluated both theoretically and by means of simulation. The results indicate that flexible margin can be selected within the range of $(1.575 \sigma_R, 2.025 \sigma_R)$ for achieving reasonable statistical assurance, for example, controlling type I error at the $\alpha=5\%$ level of significance and 90% power for analytical similarity assessment.

Keywords: Mean Shift; Inflation of Variability; Sample Size Requirement

AUTHORS/INSTITUTIONS: M. Oh, S. Lee, Celltrion Inc., Incheon, KOREA (THE REPUBLIC OF)|S. Chow, Duke University, Durham, North Carolina, UNITED STATES|
Abstract Body: Numerous statistical methods have been developed for analyzing high-dimensional data. These methods often focus on variable selection approaches but are limited for the purpose of testing with high-dimensional data. They are often required to have explicit likelihood functions. In this paper, we propose a "hybrid omnibus test" for high-dimensional data testing purpose with much weaker requirements. Our hybrid omnibus test is developed under a semiparametric framework where a likelihood function is no longer necessary. Our test is a version of a frequentest-Bayesian hybrid score-type test for a generalized partially linear single index model, which has a link function being (a functional) of a set of variables through a generalized partially linear single index. We propose an efficient score based on estimating equations, define local tests, and then construct our hybrid omnibus test using local tests. We compare our approach with an empirical likelihood ratio test and Bayesian inference based on Bayes factors, using simulation studies. Our simulation results suggest that our approach outperforms the others, in terms of Type I error, power, and computational cost in both the low-and high-dimensional cases. The advantage of our approach is demonstrated by applying it to genetic pathway data for type II diabetes mellitus.

AUTHORS/INSTITUTIONS: I. Kim, Y. Xu, Statistics, Virginia Tech, Blacksburg, Virginia, UNITED STATES|R. Carroll, Statistics, Texas A&M University, College Station, Texas, UNITED STATES|
Regression is by far the most used tool to explain relationships among variables and it is a powerful technique. However, classical linear regression estimators (OLS) are known to fail in the presence of outliers due to their extreme sensitivity. New robust estimators have been proposed in the literature especially for models with non-Gaussian errors. Koenker and Bassett proposed the quantile regression, allowing both the location and the spread of the response variable. Quantiles are computed by optimizing an asymmetrically weighted L1 norm, i.e. the sum of absolute values of residuals. Newey and Powell underlining important drawbacks to the quantiles approach, thereby they proposed an efficient approach based on the least asymmetrically weighted squares as an alternative to quantile. Furthermore, Schnabel and Eilers proposed a model combining LAWS with P-spline enhancing the flexibility of each p-expectile, leading to a higher fit for the observed trend. In some subfield of the population investigation as well as among health and medical clinical studies is common to deal with data that show the high discontinuous spread of the response to be studied. Canonical methods could not be appropriate, because the quantile (or expectile) estimations became highly influenced by chunk with high spread, jeopardizing the part where the behavior is linear. This happens in the case of life expectancy at birth or about aggregate mortality rates at older ages. The principal feature of this data is that they present a linear (or discontinued linear) shape in the first part of the data, in the final part they show a high spread level. Since the first and more regular part, could be modeled by using simple regression, the second part is not straightforward, then it is common practice to cut it, treating it separately, by using a quantile approach, as if it was a new set of data. In the same cases, as happening for life expectancy or mortality rates, it should be important to consider the continuous variation among data, thus data cutting is not the best practice.

Our paper contributes to literature proposing a novel approach based on a segmented regression pattern jointly with an expectile regression, in order to be able to model this kind of data. In this way, we are able to provide a complete overview of the changes among data in the whole data set, both in the linear part as well as in the part that shows high spread levels.

Authors/Institutions: A. Nigri, Statistics, Sapienza University of Rome, Foggia, ITALY
There is an increasing debate in the literature related to clinical trials designed upon the assumption of proportional hazards.

Recently, in cancer clinical trials for immunotherapy treatments, distinct deviations from proportional hazards are observed, rendering relevant the question of whether designing and powering a study based on this assumption leads to valid results. It is known that if the assumption does not hold, log-rank based tests and designs derived from them, may exhibit poor power and hence need to improve. The literature contains proposals based on designing the trial using other test statistics or other underlying distributional assumptions.

In this paper we present novel adaptive designs for survival outcomes for phase III trials. The central idea is to design the trial with two stages. At the end of the first stage a decision is made whether to continue but also whether the proportionality assumption is a reasonable one and if not the sample size of the second stage is re-estimated using other test statistics. In particular we make use of the restricted mean survival time (RMST) which also allows to investigate the maturity of the data in the sense that we have sufficient information to decide on the proportionality assumption. Using in advance tests for non-proportionality to power the study could also impact the sample size and power required.

Simulation results demonstrate the gain in the number of patients using this adaptive approach as well as the power of such an approach.

Extensions in more complicated designs will be also discussed.

**AUTHORS/INSTITUTIONS:** D. Karlis, C. Stavropoulos, Statistics, Athens University of Economics, Athens, GREECE|U. Dafni, Nursing, National and Kapodistrian University of Athens, Athens, GREECE|U. Dafni, Frontier Science Foundation Hellas, Athens, GREECE
Abstract Body: The receiver operating characteristic (ROC) curves and the Brier score (BS) are two common metrics used to evaluate existing prediction models of a binary outcome (Y), such as using biomarkers (X) to predict the risk of developing a disease in the future. Validation of an existing model on an external data set is an important step in determining the usefulness of the model. We consider the setting where the external dataset has some missing values for some of the covariates. We consider multiple imputation (MI) and inverse probability weighted (IPW) approaches and propose a new double-robust augmented inverse probability weighted (AIPW) estimate of the area under the ROC curve (AUC) and the BS. MI requires a model for the distribution of the variables that have missing values. IPW requires a model for the probability of missingness, while AIPW requires a model for both. A question of practical interest is whether the binary outcome variable (Y) should be included as a covariate in the models that are required to implement the approaches. We evaluated the performance of the methods in simulation studies under missing complete at random (MCAR) and missing at random (MAR). With appropriate choice of the implementation models all the methods can have very small bias. MI is more efficient than IPW. AIPW can improve the efficiency of IPW, but is not as efficient as MI in the scenarios considered. The outcome variable should be included in the model for the missing variable for MI and AIPW, while it only needs to be included in missingness model for the IPW and AIPW methods if the missingness depends on the outcome. We illustrate these methods using a prostate cancer example.

AUTHORS/INSTITUTIONS: J.M. Taylor, P. Li, M. Schipper, Biostatistics, University of Michigan, Ann Arbor, Michigan, UNITED STATES
Title: Hybrid Landmark estimation of transition probabilities

Abstract Body: Violations of the Markov assumption in multi-state models is not uncommon and problematic for estimation of many parameters of interest. The assumption is seldomly checked resulting in biased estimation. Datta and Satten (2001) showed that the Aalen-Johansen estimator of occupation probabilities is consistent also in the non-Markov case. Putter and Spitoni (2018) exploits this fact to construct a consistent estimator of state transition probabilities, the so-called landmark Aalen-Johansen estimator, which does not rely on the Markov assumption. A consequence of landmarking is data reduction leading to a loss of power. This is problematic for “less traveled” transitions, and undesirable when such transitions indeed are Markov. Using the framework of Partially non-Markov multi-state models we suggest a hybrid landmark Aalen-Johansen estimator of transition probabilities. The proposed estimator is a compromise between Aalen-Johansen estimation and landmark estimation, using transition specific landmarking. Inspired by the Markov test of Putter and Titman (2019) we use a transition specific Markov-test as a selection criterion to obtain a pragmatic estimator utilizing a larger sample than the landmark method. The methods are applied in a three-state simulation study and to real world data on individual transitions between states of sick leave, disability, education, work and unemployment. The real world data includes information from various national registries on 52128 Norwegian men born between 1971 and 1976, followed up from July 1st the year subjects turned 21 (1992–1997) and until December 31st 14 years later. The results show that the hybrid approach can drastically improve statistical power.

Authors/Institutions: N. Maltzahn, R. Hoff, J. Gran, Oslo Center for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, NORWAY|O. Aalen, J. Gran, Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, NORWAY|H. Putter, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, NETHERLANDS|
In personalized medicine, the identification of specific genes that are decisive for a particular cancer type is crucial, but the well-known n<<p issue in high dimensional gene expression data poses a severe challenge for standard classification methods. Common approaches to address this problem are either reducing the number of genes considered to be the best genes with regard to some univariate selection criteria, based on p-values or variance based measures, or to aim at estimation and variable selection simultaneously, in particular by relying on L1-regularized generalized linear regression models via the Least Absolute Shrinkage And Selection Operator (LASSO) (Tibshirani (1996, Royal Statistical Society)). LASSO often leads to very sparse solutions, setting most of the coefficients to exactly zero. A drawback of the LASSO is its over-shrinkage, forcing important covariates to have smaller coefficients. To alleviate this, the adaptive LASSO (Zou (2012, Journal of the American Statistical Association)) allows weights with different penalties for each covariate. This reduces over-shrinkage of important variables and more efficiently shrinks away the noise variables.

We propose Target Adaptive Individual Loss LASSO (tailLASSO), which synthesizes both ideas. In a first step, weights are determined proportional to their univariate performance, where we advocate for considering different clustering evaluation metrics on each gene individually, using the target variable as a grouping factor. The better a variable decomposes into distinct clusters with respect to the target, the higher weight it will receive. This reflects the assumption, motivated by the genetic background, that decisive genes with discriminative power, express differently in each target group.

In a second step, we fit the adaptive Lasso model utilizing these group-sensitive univariate weights. This can be seen as a soft filtering as we do not exclude any covariates before performing regression. Instead, it gives potentially unimportant covariates a higher penalty, while reducing the penalty on the important covariates, mitigating the over-shrinkage.

Applying our method to five well-known microarray benchmark datasets shows promising results: tailLASSO leads in average to sparser solutions while outperforming classical LASSO in terms of prediction accuracy.

**Authors/Institutions:** T. Augustin, Statistics, LMU Munich, Munich, GERMANY|C. Fütterer, M. Nalenz, Statistics, LMU Munich, Munich, GERMANY|
Abstract Body: We consider generalised ridge regression in clinical prediction settings, in particular binary and survival, for high-dimensional data. We use complementary data ("co-data", e.g. related studies, genomic annotation or cell line data) to define possibly overlapping or hierarchical covariate groups (e.g. gene sets, known signatures, Gene Ontology trees) that may differ considerably in terms of predictive strength. If so, penalising these groups by different ridge penalties likely improves prediction.

We present an Empirical Bayes approach to estimate the group penalties. Here, we provide an extra level of shrinkage to obtain stable group parameter estimates and to account for structure of the co-data. Any type of shrinkage can be used at this level, rendering a new, flexible framework to improve predictions. Moreover, the framework allows for integration and weighting of multiple co-data sets, plus posterior variable selection.

We demonstrate the method on an application to cancer genomics, in which we combine various sources of co-data and shrinkage types of the group parameters. Besides, we compare predictive performance with other commonly used methods, such as group lasso, which account for one single grouping structure, but which are not able to both shrink and estimate multiple group penalties from multiple sources. We show that the multi-group penalties stabilise variable selection, and improve the performance of parsimonious prognostic models.

**Authors/Institutions:** M. van Nee, M. van de Wiel, Epidemiology & Biostatistics, Amsterdam University Medical Centers, Amsterdam, NETHERLANDS|M. van de Wiel, MRC Biostatistics Unit, Cambridge University, Cambridge, UNITED KINGDOM|
Abstract Body: In recent years, there has been an increasing interest in estimating causal effects from observational data in epidemiology, and a wide array of sophisticated statistical methods has been developed. However, most of these methods are confirmatory by nature. This limits the scope of causal inquiries that can be made, as the methods are prone to address unsurprising hypotheses that do not question our existing view of the world. This may result in bias unintentionally being reproduced.

A natural suggestion to overcome these problems in the era of data science is to use data driven methods such as machine learning and artificial intelligence. Causal discovery is a family of statistical methods that may be considered exactly such a data-driven, exploratory alternative to classical model-driven causal inference methodology, as they seek to infer causal models directly from data. While some of the causal discovery methods have been available since the early 90s, their use in empirical research is limited to only a few studies.

In this presentation, we aim to help bridge this gap between theory and practice. We propose a new pipeline for generating causal hypotheses in life course epidemiology by use of causal discovery. This includes dividing the life course into discrete and disjoint periods, and performing causal discovery under the restriction that no causal links can move against the direction of time between such periods. Thereby, we produce a candidate causal model from a given dataset that can afterwards be perused for interesting and relevant causal hypotheses. Along with the methodological pipeline, we provide a guideline for choosing which specific causal discovery procedure to apply for different data types.

We showcase the proposed methodology on data from the Metropolit Cohort, encompassing almost 3000 Danish men born in 1953, followed from birth until the age of 65. Using this data example, we generate causal hypotheses concerning the role of socio-economic and health-related factors throughout the life course on development of depression and alcohol abuse in early elderdom.

AUTHORS/INSTITUTIONS: M. Osler, Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospitals, Copenhagen, DENMARK|A. Petersen, C. Ekstrom, Section of Biostatistics, University of Copenhagen, Copenhagen, DENMARK|
Abstract Body: A statement often seen in the news concerning some public health outcome is that a trend has changed or been broken. Such statements are often based on longitudinal data from e.g., population surveys, and the change in trend is claimed to have occurred at the time of the last data collection. These types of statistical assessments are very important as they may potentially influence public health decisions on a national level.

To assist such decision-making processes we propose two measures for quantifying the trendiness of a trend based on historical data. Under the assumption that reality evolves in continuous time we define what constitutes a trend and a change in a trend, and we introduce a probabilistic Trend Direction Index (TDI). This index has the intuitive interpretation of being the probability that a latent characteristic has changed monotonicity at any given time conditional on observed data. We also define a global index of Expected Trend Instability (ETI) quantifying the expected number of times that a trend has changed on a time interval.

Using a latent Gaussian process model we show how the Trend Direction Index and the Expected Trend Instability can be estimated in a Bayesian framework implemented in Stan, and give an application to the evolution of the proportion of smokers in Denmark during the last 20 years using data published by The Danish Health Authority.

AUTHORS/INSTITUTIONS: A. Jensen, C. Ekstrom, Biostatistics, Institute of Public Health, University of Copenhagen, Copenhagen, DENMARK
In many clinical studies, it has become increasingly common to record the values of longitudinal outcomes until the occurrence of an event of interest. With the development of electronic health records and an increased focus on personalized medicine, the need to implement multivariate models that account for a large number of longitudinal outcomes simultaneously is critical. Despite this, the state-of-the-art methods in this context have predominantly focused on univariate data, or on problems where the number of longitudinal outcomes are very low.

Our paper introduces a prognostic method called lights (generaLized joInt hiGH-dimensional longiTudinal Survival) to deal with the problem of joint modeling of longitudinal data and censored durations, in a high-dimensional context. The later introduces a latent variable modeling the heterogeneity within the patient population, with subgroups at different risks, and supposes a group-specific marker trajectory with a generalized linear mixed model for each longitudinal marker given the subgroup. Furthermore, a group-specific Cox risk of event includes multiple shared associations defined through a known functional family.

Inference is achieved using a novel fast stochastic approximation of a quasi-newton EM algorithm, by minimizing the negative log-likelihood penalized with elastic-net or group lasso regularization on the different parameter vectors of the model, depending on the desired interpretability power.

The estimated latent class membership posterior probabilities are used as discriminative marker rule in the cross-validation procedure for selecting the best regularization hyper-parameters.

The statistical performance of the method is examined on an extensive Monte Carlo simulation study, and finally illustrated on a publicly available dataset.

On this high-dimensional dataset, our proposed method is compared to the state-of-the-art models regarding risk prediction in terms of C-index, and regarding computing times. It provides powerful interpretability by automatically pinpointing significant covariates being relevant from a clinical perspective. Thus, we propose a powerful tool for personalized medicine, with the ability of automatically determining significant prognostic longitudinal biomarkers, which is of increasing importance in many areas of medicine.
Title: Bayesian optimality of testing procedures for survival data

Abstract Body: Most statistical tests for treatment effects used in randomized clinical trials with survival outcomes are based on the proportional hazards assumption, which often fails in practice. Data from early exploratory studies may provide evidence of non-proportional hazards which can guide the choice of alternative tests in the design of practice-changing confirmatory trials. We study a test to detect treatment effects in a late-stage trial which accounts for the deviations from proportional hazards suggested by early-stage data. Conditional on early-stage data, among all tests which control the frequentist Type I error rate at a fixed level, our testing procedure maximizes the Bayesian prediction of the finite-sample power. Hence, the proposed test provides a useful benchmark for other tests commonly used in presence of non-proportional hazards, for example weighted log-rank tests. We illustrate the approach in a simulation study based on data from a published cancer immunotherapy phase III trial.

Authors/Institutions: A. Arfè, Bocconi University, Milan, Italy | A. Arfè, B. Alexander, L. Trippa, Harvard Medical School, Boston, Massachusetts, United States |
Pliable lasso model (Tibshirani and Friedman 2019) is a type of the lasso (Tibshirani 1996) that considers sparsity of the interaction effects as well as main effects when high-dimensional data is considered in the linear model. In a hierarchical fashion, the model added a weak hierarchy constraint that the interaction effects are zero only if the main effects are zero. In this paper, we propose an extended pliable lasso that considers grouped covariates in the linear regression. $l_2$ penalties for the grouped variables are added in the objective function for the pliable lasso, and our proposed algorithm to solve the objective function is conducted. We first check the sparsity of the main and interaction effects simultaneously corresponding to a group. If there is at least one nonzero effect in the group, we check the sparsity of the main and interaction effects corresponding to each covariate within the group. To demonstrate the performance of our proposed method, we conduct simulation studies by comparing it with the pliable lasso and the Lasso. In a real data analysis, we use a dataset based on the Irritable Bowel Syndrome (IBS) study which is a functional gut disorder that typically manifests in early adult years. There are 80 patients and 152 variables. An interference pain variable is used as the response variable and some of the variables are used as modifying variables (gender, education, ethnicity, ...) to generate the interaction terms. We consider the remaining variables as covariates. To apply our proposed method, the grouped covariates are needed. There are largely three different types of covariates; questionnaire, single nucleotide polymorphism (SNPs) genotyping, and gut microbiome sequencing. For the variables from the questionnaire, we can group them as PROMIS variables (anxiety, cognition, ...), coping strategy variables (praying, catastrophizing, ...), Food intake-related variables (alcohol, carbohydrate, ...), and Quantitative sensory testing variables (MDT, MPS, ...). The SNPs genotyping includes oxytocin receptor gene (OXTR), glucocorticoid receptor gene (NR3C1), OPRM1, COMT, and CYP2D6. E. Each gene can be considered as a grouped categorical variable. For the gut microbiome sequencing, we group them by the phylum which is a level of taxonomic rank (Bacteroidetes, Firmicutes, ...). We chose 56 observations randomly to fit the model and use the remaining 24 observations as a test set.

AUTHORS/INSTITUTIONS: J. Lee, Statistics, University of Connecticut, Ellington, Connecticut, UNITED STATES
TITLE: Estimation of a buffering window in functional linear cox regression models for spatially-defined environmental exposure

ABSTRACT BODY: In environmental health research, it is of interest in understanding the effect of the neighborhood environment on health. Typically, neighborhood environmental exposures are measured within radial buffer zones from a residential address and identification of a buffer window is of importance, which is so called “uncertain geographic context” problem. We propose to address geographic uncertainty through developing methods for estimating the buffering window in a functional linear Cox proportional hazard model. The theoretical properties of our proposed method are studied and simulation studies are conducted. The method is illustrated in a study of the effect of walkability on cardiovascular disease in the Nurses’ Health Study.

AUTHORS/INSTITUTIONS: J. Lee, M. Wang, Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, UNITED STATES|D. Spiegelman, Biostatistics, Yale school of public health, New Haven, Connecticut, UNITED STATES|
CONTROL ID: 3376717
TITLE: Clustering of Intervals and Histogram Data

ABSTRACT BODY:

Abstract Body: The concept of symbolic data originates in Diday (1987). We consider two aspects of cluster methodology. First, while there has been a lot of activity in using regression-based algorithms to partition a data set into clusters for classical data, no such algorithms have been developed for a set of interval-valued observations. A new algorithm is proposed based on the k-means algorithm of MacQueen (1967) and the dynamical partitioning method of Diday (1973) and Diday and Simon (1976), with the partitioning criteria being based on establishing regression models for each sub-cluster. Second, we extend the Kim (2009) and Brito and Chavent (2012) (both of which extended the Chavent, 1998, work on intervals) divisive clustering for histograms based on the data midpoints to a double divisive monothetic method based on both the histogram means and their variances; see Kim and Billard (2018).

AUTHORS/INSTITUTIONS: L. Billard, Department of Statistics, University of Georgia, Athens, Georgia, UNITED STATES
Abstract Body: Nearest neighbor methods based on first differences are an approach to spatial analysis of field trials with a long history, going back to the early work by Papadakis first published in 1937. These methods are closely related to a geostatistical model that assumes spatial covariance to be a linear function of distance. Recently, P-splines have been proposed as a flexible alternative to spatial analysis of field trials. On the surface, P-splines may appear like a completely new type of method, but closer scrutiny reveals intimate ties with earlier proposals based on first differences and the linear variance model. This paper studies these relations in detail, first focussing on one-dimensional spatial models and then extending to the two-dimensional case. Two yield trial datasets serve to illustrate the methods and their equivalence relations. Parsimonious linear variance and random walk models are suggested as a good point of departure for exploring possible improvements of model-fit via the flexible P-spline framework.

Authors/Institutions: H. Piepho, University of Hohenheim, Stuttgart, GERMANY| M. Boer, Biometris, Wageningen University and Research, Wageningen, NETHERLANDS| E. Williams, Statistical Consulting Unit, Australian National University, Canberra, Australian Capital Territory, AUSTRALIA|
Causal effect estimation for competing risk data in randomized trial: adjusting high-dimensional covariates to gain efficiency

Abstract Body: The double blinded randomization trial is considered as the gold standard to estimate the average causal effect (ACE). It is known that the crude estimator without adjusting any covariate is consistent. However, in most cases, incorporating the information of covariates that are strong predictors of the outcome by adjusting them could reduce the issue of unbalance covariate distribution between treated and controlled group and can improve efficiency. Although the adjusted estimator for ACE and its properties have been well studied with low-dimensional setting with pre-specified parametric form of the covariate effect, it is unknown whether the same results hold when the dimension of covariates increase with the sample size and the form of the covariate effect is unknown. Recent work has shown that thanks to the randomization, for linear regression, an estimator under risk consistency (e.g., Random Forest) for the regression coefficients could maintain the $n^{1/2}$ convergence rate even when nonparametric model are assumed for the effect of high dimensional covariates. Also, such adjusted estimator will always lead to efficiency gain comparing to the crude unadjusted estimator. In this paper, we extend this result to the competing risk data setting and showed that under similar assumptions, the augmented inverse probability censoring weighting (AIPCW) based adjusted estimator has the same $n^{1/2}$ convergence rate and efficiency gain. Extensive simulation was performed to show the efficiency gain in the finite sample setting and a mimic clinical trial was presented to illustrate our method with adjusting for high-dimensional cytogenetic abnormality.

Authors/Institutions: Y. Cho, Mathematical Sciences, The University of Texas at El Paso, El Paso, Texas, UNITED STATES|C. Zheng, Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, UNITED STATES|M. Zhang, Biostatistics, Medical College of Wisconsin, Milwaukee, Wisconsin, UNITED STATES
Abstract:
Exposure measurement error causes bias in estimated exposure-outcome relationships and loss of statistical power. The effect of nondifferential measurement error in continuous covariates is rather well known, and techniques for adjusting for measurement error, such as regression calibration approach, have become rather popular in epidemiologic research. Epidemiologists often categorize continuous covariates into quantiles and fit regression models to them using a linear predictor on an appropriate scale. Categorization of covariates measured with nondifferential error generally leads to differential misclassification which may bias the estimated regression parameters in any direction and preclude direct use of regression calibration to reduce this bias. Under the assumption that, in the continuous scale, the true regression model is well approximated by a generalized linear model, perhaps after an appropriate covariate transformation, we provide a relatively simple approximation to the effect of the induced misclassification and adjustment for it. For a simple univariate exposure-outcome model, it is shown that despite differential misclassification, the log RR between any exposure quantiles is always attenuated with the attenuation factor equal to the correlation coefficient between true and measured exposure on the continuous scale. For a multivariate model the developed methodology is demonstrated using simulations.

Authors: V. Kipnis, Biometry Research Group, USA National Cancer Institute, Bethesda, Maryland, UNITED STATES
Penalized Cox regression with Elastic Net penalty (EN) is effective to solve the problem of prognostic factor screening and prediction model building for high-dimensional data. However, experimental or recoding error, sample heterogeneity, cause outliers in the data. These outliers may distort the estimation of EN. If these outliers are not the result of the experimental or recoding errors, that probably means the survival times of these patients relative to its covariates have different response patterns. Identification and analysis of these outliers are likely for us to find new prognosis factor and individualized treatment for them. Outliers are challenging to be detected if they located near to each other, called masking. Robust methods can solve this problem well through trimming. In this article, a penalized maximum trimmed likelihood estimator for EN-type penalized Cox regression is proposed (MPTL-COX). An improved concentration step algorithm was adopted to find the solution of MPTL-COX). Simulation studies illustrated that robust MPTL-COX performed better than EN-type penalized Cox regression in variables selection, outliers detection and prediction in high-dimensional datasets with outliers. A gene expression data from glioma patients were analyzed to illustrate its application.
Contextual multi-armed bandit (MAB) algorithms have been shown promising for maximizing cumulative rewards in sequential decision tasks such as mobile health systems. Most of the existing algorithms require the dimension of the context not be too large. This is because the cumulative gap between the rewards of the optimal arm and the chosen arm, namely regret, is shown to be proportional to a polynomial function of the dimension of the context, d. In modern applications however, it is often the case that the web or mobile-based contextual variables are high-dimensional, with only a sparse subset of size s (≪d) being correlated with the reward. We consider the stochastic linear contextual bandit problem and propose a novel algorithm, namely the Doubly-Robust Lasso Bandit algorithm, which exploits the sparse structure of the regression parameter as in Lasso, while blending the doubly-robust technique used in missing data literature. The high-probability upper bound of the regret incurred by the proposed algorithm does not depend on the number of arms and scales with the logarithm of d instead of a polynomial function of d. The proposed algorithm shows good performance when contexts of different arms are correlated and requires less tuning parameters than existing methods.

AUTHORS/INSTITUTIONS: G. Kim, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|M. Paik, Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)
Abstract Body: Background: The worldwide AIDS pandemic is caused by the human immunodeficiency virus type 1 (HIV-1). Without treatment, HIV-1 infection is almost always fatal. Another virus, HIV-2, has predominantly spread in West Africa. HIV-2 infected individuals do not always show disease progression. The purpose of our study was i) to estimate the probability of having disease progression after HIV-2 infection and ii) to compare mortality in HIV-2 progressors and HIV-1 infected individuals. We used data from a prevalent cohort study in The Gambia.

Methods: We quantified progression via a hidden Markov model, with states determined by CD4 count, and death as absorbing state. We allowed for "cure", i.e. non-progression, which was defined as having a stable CD4 count. Our program consists of two parts. In the first part, we fitted individual CD4 trajectories via a linear random effects model. At each MCMC iteration, individuals were assigned to one of the statuses (progressor or non-progressor) with a probability based on their CD4 development. Individuals classified as progressor were included in the Markov model. Transition times between CD4 states were interval censored; CD4 states at the observation times were determined by the fitted CD4 counts. We implemented our model in WinBUGS, OpenBUGS and JAGS. In WinBUGS/OpenBUGS the Markov model is specified via 42 Chapman-Kolmogorov differential equations; in JAGS we used the matrix exponential module to relate transition probabilities to transition rates. We didn't allow information to flow back to the first part. In WinBUGS/OpenBUGS this is implemented via the cut function; in JAGS we used a multiple imputation approach.

Results: While results from WinBUGS and JAGS were similar, OpenBUGS gave very different results. This is probably due to problems with implementation of the cut function (Plummer, 2015). WinBUGS and JAGS estimated the probability to be a non-progressor at 29% for HIV-2 and only 0.5% for HIV-1 infected individuals. After exclusion of non-progressors, HIV-2 infected individuals still had slower progression to death.

Discussion: Our code gives a framework for fitting hidden Markov models with interval censored transitions. JAGS allows for the most elegant implementation. We advise not to use the OpenBUGS implementation of the cut function.
Abstract Body: The typical output of Mendelian cancer risk models such as Claus-Easton (Claus, 1991) or MMRpro (Chen, 2006), BODICEA (Antoniou et al., 2008; Lee et al. 2014) is the marginal posterior carrier probability of an individual of interest, and hence his/her resulting tumoral risk. These quantities are insufficient when considering a familial risk and/or joint distributions among a set of individuals (e.g. alive relatives, siblings, a given nuclear family, etc.). In this work, we propose a method to compute the distribution of the number of carriers in a family as well as carrier distributions and joint carrier distributions conditional on this number.

Mendelian cancer risk models are classically defined as Bayesian networks (Koller, 2009) where the summation over all the unknown genotypic configurations is performed through the sum-product algorithm (Elston-Stewart, 1980; Totir, 2009; Koller, 2009). In this work we introduce an algebraic extension of the sum-product algorithm by introducing polynomials of dummy variables in order to compute the probability generating function (pgf) and the moment generating function (mgf) of the number N of carriers among the family or targeted individuals.

From the pgf of N, one can derive: \( P(N>0) \) the probability of a group to contain a carrier, \( P(N=k) \) the probability of having exactly k carriers among the targeted individuals, or all marginal carrier probabilities conditional on \( N=k \). On several example pedigrees (breast/ovarian and MSI families) we illustrate the interest of these quantities compared to simple marginal distributions for pointing out potential carriers.

By using our modified version of the sum-product algorithm, it is possible to compute extended posterior carrier distributions in a group of individuals of interest in order to help prioritizing genetic investigations. These computations can be adapted to any Mendelian cancer risk model and should provide useful information to practitioners.

The idea of using the closed testing procedure for multi-group comparisons goes back to the seventies of the last century [1]. The closed testing procedure tests null hypotheses from a set \( W \) that is closed under intersection at the unadjusted level of significance \( \alpha \). A \( H_0 \) in \( W \) is rejected if and only if all \( H_0' \) in \( W \) with \( H_0' \) implies \( H_0' \) are rejected at level \( \alpha \). The closed testing procedure controls the familywise alpha error probability by \( \alpha \). If \( k \) groups are to be compared with the expected values \( \mu_1, \ldots, \mu_k \), all null hypotheses that can be generated by intersection from the null hypotheses \( H_0: \mu_i = \mu_j, 1 \leq i < j \leq k \) must be tested. The testing of these null hypotheses is usually done by variance analytical methods (ANOVA). However, it is known that under certain alternative hypotheses the power of ANOVA is lower than that of tests based on the maximum of T-statistics (c.f. [2]). We compare the ANOVA-based closed testing procedure with Tukey's post-hoc test that is based on the maximum range of T-statistics. We show that for a wide range of alternative hypotheses the closed testing procedure is superior to Tukey's post-hoc test.

A major limitation of the closed testing procedure is its computational effort. For example comparisons of \( k=10 \) groups require the performance of 115,974 ANOVA tests. We show that by clever programming the computational time for not too large \( k \) can be reduced to a tolerable level and thus the closed testing procedure can be used in many situations relevant in practice.


AUTHORS/INSTITUTIONS: K. Neumann, Institute of Biometry and Clinical Epidemiology, Charité, University Medicine Berlin, Berlin, GERMANY
Managing chronic hepatitis C (CHC) is difficult because the majority of those infected have clinically silent disease. The asymptomatic nature of CHC means that the disease often remains undiagnosed, leaving its prevalence highly uncertain. This generates significant uncertainty for policymakers in the planning of hepatitis C eradication programs that aim to meet the goals set by the WHO. The objective of this work is to establish a mathematical framework for the estimation of CHC prevalence and the undiagnosed proportion.

Methods
A state-transition model describing infection, disease progression and treatment response was mathematically formulated and developed. Clinical health states such as cirrhosis, hepatocellular carcinoma (HCC), and decompensated cirrhosis were captured. Model parameters were obtained from the published literature. We then back-calculated the historical prevalence of CHC through a calibration process based on a Bayesian Markov chain Monte Carlo (MCMC) algorithm. The algorithm constructs posterior distributions of the historical prevalence of CHC by comparing the model-generated predictions of the annual numbers of health events related to CHC infection and its sequelae against observed calibration targets. Calibration targets include the annual numbers of HCC and HCV diagnoses recorded over a period of several years.

Results
The mathematical framework was used to estimate the CHC population among three Canadian birth cohorts. The prevalence of CHC in 2013, averaged over all birth cohorts, was estimated to be 0.63% (95% CI: 0.53% - 0.72%). The percentage of CHC cases undiagnosed in 2013 was estimated to be 27.1% (95% CI: 19.3% - 36.1%). Among the ‘baby boomer’ generation (birth cohort 1945-1964), CHC prevalence was estimated at 1.13% (95% CI: 0.92% - 1.33%), with 18.8% (95% CI: 13.3% - 23.9%) of individuals with CHC undiagnosed. Our results are in-line with earlier studies, including a recently conducted seroprevalence survey.

Discussion
Prevalence estimates impact decision making with respect to HCV public health interventions such as the introduction of screening and treatment programmes. Recent advances in the treatment of CHC will thus necessitate the updating of HCV prevalence estimates. Our study provides a platform for estimating these statistics in a robust and efficient way.

AUTHORS/INSTITUTIONS: A. Hamadeh, W. Wong, School of Pharmacy, University of Waterloo, Kitchener, Ontario, CANADA|Z. Feng, Mathematics and Statistics, University of Guelph, Guelph, Ontario, CANADA|
Abstract Body: While it is well known that high levels of prenatal alcohol exposure (PAE) result in serious cognitive and behavioural deficits in children, the exact nature of the dose response is less well understood. In particular, there is a pressing need to identify the levels of PAE associated with an increased risk of clinically significant adverse effects. To address this issue, data have been combined from six birth cohort studies in the United States which assessed PAE along with developmental outcomes measured throughout childhood. We argue that structural equation models (SEMs) provide the most appropriate way to capture the association between multiple observed outcomes, as well as to characterise the underlying variable of interest, cognition, and then relate this to PAE. Unfortunately, classic SEM software does not work well in our context, due to the variation in the outcomes being measured among the six studies as well as the complex nature of the variables. We show how a Bayesian approach can be used to fit a multilevel structural model that maps cognition to a broad range of observed variables that in turn map to several different subdomains and which also may be measured at multiple ages. The model adjusts for confounding through propensity scores. In comparison to more commonly used frequentist approaches, the Bayesian model allows us to incorporate expert knowledge into the model via weakly informative prior distributions. This approach is shown to have more stable performance and produce more reliable estimates than frequentist approaches.

Authors/Institutions: K. Dang, L. Ryan, School of Mathematical and Physical Sciences, University of Technology Sydney, Sydney, New South Wales, AUSTRALIA|K. Dang, L. Ryan, ARC Centre of Excellence for Mathematical & Statistical Frontiers, Sydney, New South Wales, AUSTRALIA|T. Akkaya-Hocagil, Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Ontario, CANADA|J. Jacobson, S. Jacobson, Department of Psychiatry and Behavioral Neurosciences, Wayne State University, Detroit, Michigan, UNITED STATES|
Abstract Body: The hazard ratio (HR) is by far the most frequently used effect measure in applied time-to-event analyses. However, the causal interpretation of the HR is not necessarily simple, even in the context of randomized controlled trials (RCTs). Even if randomization ensures treatment groups that are balanced for all observed and unobserved factors at baseline, these groups would become increasingly unbalanced as time passes. The reason is that if one treatment is effective, then it will keep individuals from having the event of interest for a longer time period than if these same individuals were not given the effective treatment. Hence, more vulnerable, or frail, individuals will be in the at-risk population in the group receiving the effective treatment than in the group receiving a less effective treatment, at any time point after the first event have occurred in the study population. The hazard rate by definition conditions on being event-free up to each time point. Thus, the groups will no longer be balanced after the first event has occurred.

Recently, such selection effects have received some attention, and have been used to explain decreasing HRs in RCTs. However, the same type of selection occurs also when the observed HR is constant, i.e. when all assumptions of Cox' proportional hazards model are satisfied. Taking a frailty modelling point of view, we show that the observed, time-constant HR will be an attenuated version of the treatment effect in individuals also in this situation.

In a competing risks setting, the causal interpretation of the observed (time-constant) cause-specific HRs is even more challenging. If the same unobserved factors affect the different competing risks, then we show that the observed cause-specific HRs can be both overestimating or underestimating the effect of a treatment/exposure. Thus, treatments that are harmful or beneficial, respectively, may appear as having the opposite effect if the observed cause-specific HRs are used as the measures of effect. Hence, this renders the causal interpretation of such observed cause-specific HRs difficult.

In conclusion, making causal conclusions based on HRs are challenging, even in RCTs and even when all assumptions of the Cox model are satisfied.

Authors/Institutions: M. Valberg, M. Brathovde, Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, NORWAY| O. Aalen, Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, NORWAY|
Abstract Body: Optimizing field testing resources is a key component of agricultural evaluations, especially for large experiments. Testing efficiency can be improved by controlling micro- (i.e. spatial) and macro-environmental (i.e. time and space) variability. Spatial variability can be controlled with both, efficient experimental designs and spatial modeling. On the other hand, genotype by environment interaction can be modeled when multi-environment trials (METs) are used. However, how to allocate resources within and among environments is not entirely clear. Furthermore, because molecular markers can be used to estimate relationship among genotypes, information can be borrowed from relatives relieving the historical burden on balanced evaluations. The objective of this study was to compare strategies for micro and macro-environmental variability control that include spatial and genotype by environment modeling to optimize resource allocation in METs. Real field variability as well as real genotypic means and genotype by environment interaction structures were used with a simple simulation strategy to evaluate the performance of experimental designs, spatial corrections, and level of within and among environment replication under several plot size, heritability, and genotype by environment structures. Completely randomized design, randomized complete block designs, alpha-designs, unreplicated, partially-replicated, augmented designs, and spatial designs were compared. The use of no correlations structure on residuals, as well as AR1, AR1xAR1, and the use of splines to model residual structures was also compared. Finally, different resource allocation strategies were evaluated within and among environments to optimize response to selection. Spatial modeling is superior than assuming independent errors. However, spatial corrections are not superior than spatially minded experimental designs. Our resource allocation strategy significantly increases selection response in breeding programs when genotypes are evaluated in METs.

Authors/Institutions: L. Gutierrez, Agronomy, University of Wisconsin - Madison, Madison, Wisconsin, United States
Colorectal cancer (CRC) is a major cause of morbidity and mortality globally. It is known that CRC survival is highly dependent upon the stage of disease at diagnosis. Reductions in mortality can be achieved through the detection and treatment of early-stage CRC patients. Optical colonoscopy, which allows for direct visualization of colonic polyps, is currently the most effective CRC screening test in nowadays. However, it is costly, invasive, and requires anesthesia, which discourages many people from pursuing routine tests. A simple noninvasive test with high accuracy for CRC is urgently needed, as it might increase adherence rates, resulting in better clinical outcomes. Several recent CRC studies have demonstrated a significant association between tumorigenesis and abnormalities in the microbial community. Those findings shed light on utilizing microbial taxa as noninvasive CRC biomarkers. In this paper, we propose a Bayesian hierarchical framework to identify a set of differentially abundant taxa, which could potentially serve as microbial biomarkers. The bottom level is a multivariate count generative model that links the observed counts in each sample to their latent normalized abundances. For the choice of a zero-inflated negative binomial model as the bottom level, we use the Dirichlet process as a flexible nonparametric mixing distribution to model all unknown factors that could influence sequencing depth. The top level is a Gaussian mixture model with a feature selection scheme for identifying those taxa whose normalized abundances are discriminatory between different phenotypes. The model further employs Markov random field priors to incorporate phylogenetic tree information to identify microbial biomarkers at different taxonomic ranks. A simulation study on both simulated and synthetic data is conducted. A CRC case study demonstrates that a resulting diagnostic model trained by the microbial signatures identified by our model in a CRC cohort can significantly improve the current predictive performance in another independent CRC cohort.
Abstract: The advent of very large, richly-phenotyped and high-quality human genomics datasets, together with the development of models that allow joint analyses of all GWAS test statistics, have led to big advances in understanding the genomic architecture of complex traits. Analysis of summary statistics is effectively unlimited in sample size and number of genetic variants, and avoids confidentiality issues associated with individual-level data. However, the models rest on assumptions that were initially unchallenged but have since become the subject of controversy. A key focus has been to answer detailed questions about how much causal variation lies in specific genomic regions, particularly those with functional annotations. Different approaches have led to very different estimates, giving discordant pictures of the genomic architecture of complex traits. The problem has been different assumptions about how the causal effect of a SNP varies according to genomic properties known a priori, for example minor allele fraction and linkage disequilibrium. I will review recent progress in using genome-wide SNPs to assess the heritability of complex human traits and its distribution across the genome, as well as the effect of confounding on GWAS test statistics. We have also derived improved estimates of selection parameters, leading to new insights into the effects of purifying selection on various traits. I will describe a consensus position that we have recently reached, and our reasons for believing that we are close to robust answers about genomic architecture, even though there remain many avenues for model refinements.

Authors/Institutions: D. Balding, Melbourne Integrative Genomics, University of Melbourne, University of Melbourne, Victoria, AUSTRALIA|D. Balding, D. Speed, UCL Genetics Institute, London, UNITED KINGDOM|D. Speed, Aarhus Institute of Advanced Studies, Aarhus, DENMARK
Abstract Body: It is clear that the human's microbiome is associated with various disease. Much work has been done to analyze microbiome data to identify microbiota related with specific disease. However, in the case of study using only microbiome data, even if we identify microbiota related to a particular disease, it is difficult to know the functional potential of the microbiome. The metabolomes produced from the microbial community are known to play a role in connecting host phenotype and microbiome function. Using both metabolomics and metagenomics has an advantage of understanding functional potentials of the microbiome and interactions with the host. However, integration of these two omics data remains a challenge, usually requiring a more advanced method. In this study, we proposed the hierarchical structural component model (HisCoM-MnM) that integrates microbiome and metabolome data. In particular, we used pathway information for integrate these two omics datasets to provide insight into biological interactions between different biological layers in relation to host phenotype. We applied our model to analyze real datasets generated from specific diseases. These real datasets were used to demonstrate whether our model is able to identify the pathways known to be related with disease. This analysis shows our HisCoM-MnM can identify disease related pathways from KEGG database and provide significant metabolomic and metagenomic components of pathway.

Authors/Institutions: B. Kim, Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF) | T. Park, Department of Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)
Abstract Body: In many clinical areas, it is common to combine several clinically meaningful endpoints in a composite endpoint to evaluate the full clinical benefit of a treatment. The standard method of analysis for a composite endpoint is a time to first event analysis, which has several shortcomings. The results are often driven by lesser important components, it ignores information on subsequent events and only survival data can be combined. Hence, a time to first event analysis may still not evaluate the full treatment benefit.

Generalized pairwise comparison (GPC) analysis methods, such as the win ratio and net benefit, have been described that alleviate the shortcomings of the time to first event analysis. Although the GPC methods are increasingly being used, their properties are not well understood. We aim to evaluate the properties of the GPC methods. Early criticism mentioned the dependence of the net benefit and win ratio on the censoring distribution. We found additionally that the break-down of the treatment effect on the individual components of the composite endpoint does not represent well the true treatment effect on the component. Hence, the prioritized GPC analysis should not be used to evaluate the treatment effect on individual components. When investigating unbiasedness, sufficiency and completeness of the GPC statistics, we conclude that the win ratio statistic is not an unbiased estimator, while the net benefit statistic is. In a univariate setting with complete data, sufficiency and completeness of the net benefit statistic depends on the distribution of the data. When the distribution for the observations belongs to an unrestricted family (Bernoulli), the net benefit statistic is the uniformly minimum-variance unbiased estimator (UMVUE), while it is not when observations come from a more restrictive family of distributions (normal, exponential or Poisson). Theory and simulations show that the loss in efficiency for the net benefit statistic is limited in realistic scenarios. Since, the underlying distribution of the observations is in practice often unknown, the net benefit statistic is the safer option.

In conclusion, the limitations of the GPC statistics in the presence of censoring and on the interpretation of individual component treatment effects should be recognized. Moreover, the net benefit has better theoretical properties than the win ratio.

Abstract Body: Standard methodology for meta-analysis focuses on deriving a summary treatment effect but remains implicit about the case-mix of the patient population for which this summary statistic is described. If the individual patient data (IPD) are available from all trials of interest, the biases stemming from case-mix heterogeneity among these trials could be mitigated by using inverse probability weighting (IPW), which standardizes results of different trials over the case-mix of a target population prior to evidence synthesis (Vo et al, Research Synthesis Methods, 2019). In practice, IPD is however rarely available for all trials of interest due to the privacy and commercial concerns.

In this talk, we will discuss the feasibility of case-mix standardization when in the meta-analysis, the IPD is only accessible for some trials but not for the others. To achieve this, we first propose a novel method that allows for standardizing results of trials with IPD over the case-mix of trials without IPD. This can be done by adjusting average patient characteristics in trials with IPD to match those reported for trials without IPD. Accompanying two-random-effect meta-analysis model are developed, which enable disentangling the total heterogeneity across trials into what we term case-mix and beyond case-mix heterogeneity. The impact of each heterogeneity source on the summary estimate is then measured by two different I² statistics (or two different prediction intervals). The summarized treatment effect specific for each trial population is then derived by using the predictions of the random effect reflecting the case-mix heterogeneity in the proposed model.

In what follows, we extend the above approach to the so-called network meta-analysis. As in the case of pairwise meta-analysis, a different network of evidence will be established for each trial population, which then describes the effect of different treatments when being applied to a specific trial population defined by a proper case-mix. We evaluate the novel approach by numerically simulated data and apply it to explore the effectiveness of different Insulin-sensitising medications for women with polycystic ovary syndrome. Results of these analysis show that the new approach can lead to an insightful heterogeneity assessment in standard meta-analysis practice.

Authors/Institutions: T. Vo, Applied Mathematics, Computer Science and Statistics, Ghent university, Ghent, BELGIUM|S. Vansteelandt, Ghent University and the London School of Hygiene and Tropical Medicine, Ghent, East Flanders, BELGIUM|R. Porcher, Université de Paris, CRESS, INSERM, INRA, Paris, FRANCE
Abstract Body: Interest in the prediction of all kinds of properties, like sensory traits of foodstuff, from untargeted metabolomics data is high. To achieve the highest accuracy possible, there is a tendency to fuse data from multiple analytical platforms. However, there is no consensus on how to combine said data or what modelling techniques to use. Here, we evaluate the performance of random forests, stochastic gradient boosted machines, regularized regressors, support vector machines, partial least squares, as well as a more traditional method in the form of stepwise regression. In literature, three approaches to combine data sets are proposed, namely low, intermediate and high data fusion. An exploratory analysis of the various methods' performance across all datasets and data fusion levels is conducted. In addition, the effect of the data fusion strategies is statistically evaluated for each method and dataset. This evaluation is performed on 4 real, independent datasets. Using the Friedman, and Nemenyi post-hoc tests, we observe that regardless of method or data set, no fusion level significantly outperforms the single most informative data block. Furthermore, we note that no individual method consistently outperforms all others.

Longitudinal biomarker screening algorithms for early detection of hepatocellular carcinoma

**Abstract Body:**
Advanced hepatocellular carcinoma (HCC) has limited treatment options and poor survival. Early detection of HCC is critical to improve the prognosis of these patients. Current guidelines for high-risk patients include six-month ultrasound screenings but these are not sensitive for early HCC. Alpha-fetoprotein (AFP) is a widely used diagnostic biomarker but has shown limited use in HCC screening with a fixed threshold.

Approaches that incorporate longitudinal AFP have shown potentially increased earlier detection of HCC. A parametric empirical Bayes algorithm, first proposed by McIntosh and Urban (2003), defines a patient-specific threshold that is a weighted average of the population mean and the sample mean of the patient screening history. The PEB algorithm has been applied to HCC screening with AFP in data from the Hepatitis C Antiviral Long-term Treatment against Cirrhosis (HALT-C) trial, a randomized clinical trial, as well as a cohort constructed from the electronic medical records from the Department of Veteran’s Affairs Hepatitis C Clinical Case Registry. In both studies, we observed gains in the discriminatory performance of AFP via the PEB algorithm.

However, in most cancer settings, including HCC, a single biomarker will not cover the heterogeneous subtypes present in the target surveillance population. Des-gamma-carboxy prothrombin (DCP) is a serum biomarker that has been evaluated in Phase-2 biomarker studies and approved by the FDA for HCC risk. We have developed a multivariate parametric empirical Bayes (mPEB) algorithm that identifies optimal patient-specific thresholds for a panel of biomarkers. A minimal model for the biomarker levels in both cases and controls is specified. The optimal set of patient-specific thresholds ensures the false positive rate is at most $f_0$ while maximizing the likelihood of a positive screen if the patient has HCC, conditional on the patient screening history. The mPEB algorithm was compared to alternatives in the HALT-C Trial (using cross-validation) and in simulations studies under a variety of possible scenarios for biomarker trajectories.

**Authors/Institutions:**
N. Tayob, Data Science, Dana-Farber Cancer Institute, Boston, Massachusetts, UNITED STATES
A.S. Lok, Internal Medicine, University of Michigan, Ann Arbor, Michigan, UNITED STATES
Z. Feng, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, UNITED STATES
N. Tayob, Harvard Medical School, Boston, Massachusetts, UNITED STATES
Time-to-event analysis of genome data may enable a better understanding of the mechanisms and underlying genetic architecture behind the onset and development of common complex disease. Typically, genome datasets are very large, and the number of covariates strongly exceeds the number of observations, therefore variable selection and appropriate methods for effect estimation are required. Previous work has not properly addressed these issues when dealing with these data, potentially leading to biased effect size estimates with poor precision. Here, we propose a hierarchical Bayesian model which assumes that time-to-event has a Weibull distribution, handles sparsity with spike and slab variable selection, considers left truncation and right censoring, and also yields estimates of the total proportion of variance explained by genomic markers (SNP heritability). The model can handle large datasets by combining Gibbs sampling and adaptive rejection sampling. In simulation study, our approach outperforms previous approaches and provides better estimates of the genetic effects. We apply the approach to the real-life data on survival time and time-to-event of cardiovascular disease and stroke. Our general Bayesian framework enables more accurate discovery and estimation for the survival associated genomic marker effects, thus granting novel insight into the genetic architecture of time-to-event data.
Abstract: A new method for the analysis of time to ankylosis complication on a dataset of replanted teeth is proposed. In order to deal with these left-censored, interval-censored and right-censored data a Cox model with piecewise constant baseline hazard is introduced. Estimation is carried out with the EM algorithm by treating the true event times as unobserved variables. This estimation procedure is shown to produce a block diagonal Hessian matrix of the baseline parameters. Taking advantage of this interesting feature of the estimation method a L0 penalised likelihood method is implemented in order to automatically determine the number and locations of the cuts of the baseline hazard. This procedure allows to detect specific areas of time where patients are at greater risks for ankylosis. The method can be directly extended to the inclusion of exact observations and to a cure fraction. Theoretical results are obtained which allow to derive statistical inference of the model parameters from asymptotic likelihood theory. Through simulation studies, the penalisation technique is shown to provide a good fit of the baseline hazard and precise estimations of the resulting regression parameters.

Authors/Institutions: O. Bouaziz, MAP5 Laboratory, Université de Paris, Paris, FRANCE|G. Nuel, Institute for Mathematical Sciences, CNRS, Paris, FRANCE|E. Lauridsen, Ressource Center for Rare Oral Diseases, Copenhagen, DENMARK
Next generation sequencing technologies have made it possible to investigate the role of rare variants (RVs) in disease etiology. Because RVs associated with disease susceptibility tend to be enriched in families with affected individuals, study designs based on ascertainment of affected sib pairs (ASP) can be more powerful than conventional case-control studies of unrelated individuals. We construct tests of RV-set association in ASPs for single genomic regions as well as for multiple regions. Single-region tests can efficiently detect a genomic region harboring susceptibility variants, while multiple-region extensions are meant to capture signals dispersed across a biological pathway, potentially as a result of locus heterogeneity. Within ascertained ASPs, the test statistics contrast the frequencies of duplicate rare alleles (usually appearing on a shared haplotype) against frequencies of a single rare allele copy (appearing on non-shared haplotypes); we call these allelic parity tests. Incorporation of minor allele frequency estimates from reference populations can markedly improve test efficiency. Under various genetic penetrance models, application of these tests in simulated ASP datasets demonstrates good type I error properties as well as power gains over approaches that regress ASP rare allele counts on sharing state. The improved performance of the allelic parity tests can be explained by the fact that allele parity counting (ie. whether alleles appear as singles or duplicates) is a better discriminator between susceptibility and null regions at the sib-pair level. As proof of principle, we apply gene-based and pathway-based tests to an established DNA pathway in a dataset derived from whole exome sequencing of sisters ascertained with early onset breast cancer. Considering practical issues, we address robustness of the allelic parity methods to sequencing error, the presence of genetic linkage, population stratification, and misspecification of reference population allele frequencies.

AUTHORS/INSTITUTIONS: R. Romanescu, S.B. Bull, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, CANADA|S.B. Bull, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, CANADA|
Abstract Body: We consider the problem of estimating unknown population parameters based on data with
nonrandom missing values, focusing particularly on inverse propensity weighting and instrumental variables. With
discrete instruments, Shao and Wang (2016) proposed a semiparametric estimation of unknown population
parameters assuming an exponential tilting propensity. A naive application of this idea to continuous instrumental
variables through arbitrary discretizations is apt to be inefficient, and maybe subject to the theoretical complications of
data-dependent threshold placement. In this talk, we propose a novel and flexible approach that does not rely on
single arbitrary discretizations but involves continuous dichotomizations across the instrument space. Empirical
processes resulting from these dichotomizations are then used to estimate the unknown parameters through weighted
integration. We establish the consistency and asymptotically normality of the proposed estimator. Simulation studies
and the analysis of a real data set generated in caries research demonstrate the gains of the methodology over
procedures that rely on arbitrary discretizations.

Authors/Institutions: D. Todem, Epidemiology and Biostatistics, Michigan State University, East Lansing,
Michigan, UNITED STATES|W. Hsu, Statistics, Kansas State University, Manhattan, Kansas, UNITED STATES|
Abstract Body: More and more tools for the analysis of microbiome sequencing data are being published, but the methods used for the evaluation and benchmarking of the tools are very diverse, making a comprehensive and honest comparison very hard. We have recently argued that simulated data from parametric families of distributions (e.g. negative binomial) do not result in realistic evaluations (Hawinkel et al.). Moreover, the correlation between the taxa need to be realistically simulated as well.

We have developed a new semiparametric simulation method (NPSimSeq, Assefa et al.) that allows simulation of very realistic high dimensional datasets, while retaining realistic dependence structures and that leaves the freedom to introduce a biological signal (e.g. differential abundance). This method forms the basis of an initiative to setup a comprehensive benchmarking study for evaluating data analysis pipelines for microbiome amplicon sequencing studies. In a first stage the focus is on methods for testing for differential abundance.

We have designed an open modular architecture in the R environment that easily allows for other contributors to submit code for their data analysis method (either in R or in Python) to be evaluated and benchmarked with other methods. Reproducibility and open source are key elements in our setup. We are inviting other researchers to join our effort and to make this a collaborative community benchmarking initiative, in agreement with the guidelines recently proposed by Weber et al. (2019).

References:
Hawinkel, S., Rayner, J.C.W., Bijnens, L. and Thas, O. Sequence count data are poorly fit by the negative binomial distribution. Submitted to PLOS ONE.
Assefa, A., Vandesompele, J. and Thas, O. SPsimSeq: semi-parametric simulation of bulk and single cell RNA sequencing data. Revision submitted to Bioinformatics.
Abstract Body: In precision agriculture, spatial prediction of yield at a given site within a crop field is model-based. With the availability of precision agriculture technology, there is an increasing number of on-farm trials in which rates of agronomic inputs (treatments) are spread and highly repeated all across the space. Yield data obtained from the treatments are used to estimate site-specific functions with the aim to predict yield under different treatments and spatially distributed covariables. Careful and robust assessment of alternative methods to build spatial models that make predictions is needed. In this work, yield was modelled as a function of variable rates of nitrogen and seed from precision agriculture on-farm experiments using various methodological approaches: 1) linear regression model estimated via Restricted Maximum Likelihood (REML), 2) Bayesian regression with random site effect estimated via Integrated Nested Laplace Approximation (INLA), and 3) Random Forests (RF) as a machine learning approach. Each approach was applied with and without soil variables included as covariates. The residuals from the first and the third approaches were assumed either independent or spatially correlated. Eleven replicated and randomized field trials of variable rates for nitrogen and/or seed density, including elevation and shallow and deep electroconductivity as site covariables, were analysed using each approach. All models showed low prediction errors (<15% of yield mean), even without site covariables in the model. Spatial correlation in the error terms did not improve predictions with REML regression probably due to the large sample sizes involved in this type of experiments. The random site effects in the Bayesian regression captured spatial site variability and improved yield predictions in most of the fields. RF with kriged residuals yielded the smaller global prediction error in fields where site covariables showed high variability and spatial correlation. Dealing with spatiality sometimes does not help to improve site-specific predictions, probably due to the abundance of data in precision agriculture experiments or the presence of spatial auto-correlation captured by the mean structure of the model. The ultimate goal of an accurate spatial prediction of yield is to prescribe variable-rate inputs across the field for future crops.

Authors/Institutions: P. Paccioretti, M. Córdoba, C.I. Bruno, M. Balzarini, National Scientific and Technical Research Council, Córdoba, Córdoba, ARGENTINA
P. Paccioretti, M. Córdoba, C.I. Bruno, M. Balzarini, UNIVERSIDAD NACIONAL DE CÓRDOBA, Cordoba, Cordoba, ARGENTINA
B. David, University of Illinois, Urbana-Champaign, Illinois, UNITED STATES
Abstract Body: Statistical methodology is developed that allows peer effects to be modified by the structural importance of the focal actor's position in a social network. The methodology is first developed for a single peer effect and then extended to simultaneously model multiple peer-effects and their modifications by the structural importance of the focal actor. This work is motivated by the diffusion of implantable cardioverter defibrillators (ICDs) in patients with congestive heart failure (CHF) across a cardiovascular disease patient-sharing network of United States hospitals. We apply the general methodology to estimate peer effects for the adoption of capability to implant ICDs, the number of ICD implants performed by hospitals that are capable, and the number of patients referred to other hospitals by non-capable hospitals. Applying our novel methodology to study ICD diffusion across hospitals, we find evidence that exposure to ICD-capable peer hospitals is strongly associated with the chance a hospital becomes ICD-capable and that the direction and magnitude of the association is extensively modified by the strength of that hospital's position in the network, even after controlling for effects of geography.
Causal mediation analysis may be of interest when both the mediator and final outcome measured repeatedly, but limited work has been done for this situation. Available methods are based on parametric models and thus sensitive to model assumptions. We have developed a more flexible and robust approach to causal mediation analysis for longitudinal data via semiparametric continuous time models. Our approach uses linear mixed effects spline models for the mediator and the final outcome, which are fit in sequence, the predicted mediator from the mediator model being used as a covariate in the final outcome model. The models allow flexible functions for both the mean and individual response functions, as well as lagged effects of the mediator. An extended mediation formula and sequential ignorability assumption is employed to estimate natural direct and indirect effects, both overall and as a function of time. The method is applied to data from a cohort study to assess attention as a mediator of the effect of prenatal tobacco exposure on externalizing behavior in children.
ABSTRACT BODY:

By extending the hierarchical group-LASSO interaction network (glinternet) with an EM-style random effect estimation, we provide a new method for recovering sparse interaction signals in high-dimensional, dependent data. In this talk, the curse-of-dimensionality problem encountered by LASSO variable selection in linear mixed models including two-way interactions is studied. The ability to scale up to millions of interaction terms facilitates applications such as cross-omics systems biology, personalized medicine, crossover clinical trials, and biomarker discovery in repeated measurements.

Although the LASSO algorithm targets prediction performance, it is often deployed at the exploratory stage assuming desirable variable selection properties such as a low false negative rate (FNR) and/or low false discovery rate (FDR). An extensive simulation presents three findings on high-dimensional linear mixed models of compound symmetry type. Firstly, the LASSO tends to recover the true main effects (low FNR) at the expense of many false positives (high FDR). This behavior is magnified as the signal-to-noise ratio grows, and holds for low as well as high-dimensional problems, where the number of variables far exceeds the number of observations. Secondly, our method is moderately successful in recovering active interaction pairs among millions of candidates. For orthogonal design matrices, FNR can be substantial in high-dimensional problems. On the other hand, if the active predictors are correlated, FNR of interactions drops sharply, which may be considered surprising. Thirdly, the method estimates both variance components of the compound symmetry structure reasonably accurate, although the between subject variance component tends to be underestimated.

Based on these findings, several tools that attempt to improve the false positive / false negative tradeoff including the Relaxed LASSO, additional shrinkage, and regularization by screening are discussed.

Our approach is illustrated on a recent longitudinal dataset involving infants from HIV infected Zimbabwean mothers receiving antiretroviral treatment, identifying key contributors among many cross-omics and drug-omics interactions that predict health related outcomes over time.

AUTHORS/INSTITUTIONS: K.L. Duisters, L.B. Zwep, J.J. Meulman, Mathematical Institute, Leiden University, Leiden, NETHERLANDS|L.B. Zwep, Leiden Academic Centre for Drug Research, Leiden University, Leiden, NETHERLANDS|J.J. Meulman, Department of Statistics, Stanford University, Stanford, California, UNITED STATES|
Abstract Body: One of the important tasks for biobanks is the identification of genomic predictors for incident diseases and overall survival. As the follow-up time is still relatively short, the number of observed events can be quite small, despite of the impressive size of the overall cohort. Therefore, it has been proposed to use parental survival times as outcomes in such analyses, whereas offspring genotypes are used as covariates. In case the Cox proportional hazards model holds for the genotype-survival association, it does not hold when the offspring genotypes are used instead of true genotypes as covariates. However, assuming the proportional hazard model for the true genotypes one can derive an expression for the time-dependent hazard ratio for the offspring genotypes. Using this functional relationship it is possible to estimate the coefficients that one would have obtained had the true genotypes been available for the survival analysis.

The methodology will be illustrated using the Estonian Biobank data.

Authors/Institutions: M. Möls, M. Mändul, K. Fischer, Institute of Mathematics and Statistics, University of Tartu, Tartu, ESTONIA| M. Mändul, K. Fischer, Institute of Genomics, University of Tartu, Tartu, ESTONIA|
A comparative study on the unified model based multifactor dimensionality reduction methods for identifying gene-gene interactions associated with the survival phenotype

Abstract Body: For the gene-gene interaction analysis, the multifactor dimensionality reduction (MDR) method has been widely used to reduce multi-levels of gene-gene interactions into high or low risk groups under a single classifier. For the survival phenotype, the Cox-MDR has been proposed to use a martingale residual of a Cox model as a classifier whereas the KM-MDR has been recently proposed to use the median survival time as a classifier. As in the traditional MDR method, both Cox-MDR and KM-MDR requires the cross-validation procedure to find the best SNP pairs among all possible pairs, without any significance testing. The permutation testing should be followed to check whether the interaction of SNP pairs obtained by the cross-validation is statistically significant. However, the unified model based multifactor dimensionality reduction method (UM-MDR) has been proposed to unify the significance testing with the MDR algorithm within the regression model framework. Neither cross-validation nor permutation testing are required to identify gene-gene interactions in the UM-MDR method. Recently, the Cox UM-MDR has been proposed which combines the Cox-MDR method with the key procedure of UM-MDR to identify gene-gene interactions associated with the survival phenotype. In this paper, we propose an alternative method, called KM UM-MDR which combines the KM-MDR method with the main idea of UM-MDR. Through the simulation study, we compare the performance of KM UM-MDR with those of Cox UM-MDR, Cox-MDR and KM-MDR with and without considering the marginal effects of SNPs. We also analyze a real dataset of Korean leukemia patients to compare the proposed method with the Cox UM-MDR method.

AUTHORS/INSTITUTIONS: S. Lee, Sejong University, Seoul, KOREA (THE REPUBLIC OF); J. Lee, Statistics, University of Connecticut, Willington, Connecticut, UNITED STATES
Eide and Gefeller introduced the concepts of sequential and average attributable fractions as methods to partition the risk of disease to differing exposures. In particular, sequential attributable fractions are interpreted in terms of an incremental reduction in disease prevalence associated with removing a particular risk factor from the population, having removed other risk factors. Clearly, both concepts are causal entities, but are not usually estimated within a causal inference framework.

Here model the inter-relationships between exposures and disease using a causal Bayesian network, assuming no unobserved variables. This allows us to model not only the direct impact of removing a risk factor on disease, but also the indirect impact through the effect on the prevalence of causally downstream risk factors that are typically ignored when calculating sequential and average attributable fractions. The procedure for calculating sequential-attributable fractions involves repeated applications of the do-operator over a fitted Bayesian network, and simulation from the resulting joint probability distributions. While this procedure is computationally involved, it permits consistent estimation of joint attributable fractions which may not be possible with a single regression model for certain causal structures.

The methods are applied to the INTERSTROKE study, which was designed to quantify disease burden attributable to the major risk factors for stroke. The resulting sequential and average PAFs are compared to previous results (calculated using the R-package ‘averisk’) which used a single logistic model (adjusting for a common set of confounders) in estimating sequential PAFs.

**AUTHORS/INSTITUTIONS:** J.P. Ferguson, Clinical Research Facility, NUI Galway, Co.Galway, IRELAND
Kernel-based hierarchical structural component model with high prediction power

Abstract Body: Over the last few years, many statistical methods for pathway-based analysis have been proposed. These methods have mainly focused on association tests between pathway and phenotypes rather than on development of models with high predictive power. To develop the model with high prediction power while considering the hierarchical structural relationship between genes and pathways, we propose a new statistical approach based on kernel, “Hierarchical structural CoMponent analysis using Kernel (HisCoM-Kernel)”. HisCoM-Kernel uses individual similarities instead of original observations by introducing the kernel function to the model. Especially, we can explain the direction of pathway effect on the outcome by imposing non-negative penalty on gene effects. Through simulation studies, HisCoM-Kernel showed higher predictive power than existing comparative methods. Moreover, the application of the proposed method to real data illustrated that HisCoM-Kernel successfully identified the meaningful effect of the pathway on the phenotype compared to other methods.

AUTHORS/INSTITUTIONS: S. Hwangbo, Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|S. Lee, Center for Precision Medicine, Seoul National University Hospital, Seoul, KOREA (THE REPUBLIC OF)|I. Kim, Statistics, Virginia Tech, Blacksburg, Virginia, UNITED STATES|T. Park, Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)
Abstract Body: Use of big data is becoming increasingly popular in medical research. Since big data-based projects differ notably from classical research studies, both in terms of scope and quality, a debate is apt as to whether big data require new approaches to scientific reasoning different from those established in statistics and philosophy of science.

The progressing digitalization of our societies generates vast amounts of data that also become available for medical research. Here, the big promise of big data is to facilitate major improvements in the treatment, diagnosis and prevention of diseases. An ongoing examination of the idiosyncrasies of big data is therefore essential to ensure that the field stays congruent with the principles of evidence-based medicine. We discuss the inherent challenges and opportunities of big data in medicine from a methodological point of view, particularly highlighting the relative importance of causality and correlation in commercial and medical research settings. We make a strong case for upholding the distinction between exploratory data analysis facilitating hypothesis generation and confirmatory approaches involving hypothesis validation. An independent verification of research results will be ever more important in the context of big data, where data quality is often hampered by a lack of standardization and structuring.

Conclusions: We argue that it would be both unnecessary and dangerous to discard long-established principles of data generation, analysis and interpretation in the age of big data. While many medical research areas may reasonably benefit from big data analyses, they should nevertheless be complemented by carefully designed (prospective) studies.

Reference:
Caliebe A, Leverkus F, Antes G, Krawczak M.
Does big data require a methodological change in medical research?.

AUTHORS/INSTITUTIONS: A. Caliebe, M. Krawczak, Institute of Medical Informatics and Statistics, Kiel University, Kiel, GERMANY|A. Caliebe, M. Krawczak, University Hospital Schleswig-Holstein, Kiel, GERMANY|F. Leverkus, Pfizer Deutschland GmbH, Berlin, GERMANY|G. Antes, Institute for Medical Education, University Hospital, LMU Munich, Munich, GERMANY|
Evidence from animal models and epidemiological studies has linked prenatal alcohol exposure (PAE) to a broad range of cognitive and behavioral deficits. However, there is virtually no information in the scientific literature regarding the levels of PAE associated with an increased risk of clinically significant adverse effects. The sample size in individual prospective longitudinal cohort studies may not provide sufficient power to examine effects associated with different levels and patterns of PAE. To address this critical public health issue, we propose to synthesize information regarding the effects of PAE and cognition across multiple endpoints using data from six major U.S. longitudinal cohort studies. We propose a two-stage meta-analytic approach which involves estimating the dose-response coefficients for each endpoint, and then pooling these correlated dose-response coefficients to obtain an estimated “global” effect of exposure. Specifically, in the first stage, we used individual participant data to derive the estimates of the effect of alcohol exposure by fitting a regression models which address confounding variables through adjusting for the propensity score. The correlation matrix characterizing the dependence between the endpoint-specific dose-response coefficients estimated within each cohort is then estimated, while accommodating incomplete information on some endpoints. We then discuss and compare inferences based on the two-stage approach to inferences based on a full multivariate analysis.
Abstract: Colocalization analysis is a popular method for the quantitative analysis in fluorescence microscopy imaging. The localization of marked proteins in the cell nucleus allows a deep insight into biological processes in the nucleus. Several metrics have been developed for measuring the co-localization of two markers, however, they depend on subjective thresholding of background and the assumption of linearity.

We propose a robust method in order to estimate the bivariate distribution function of two color channels. From this, we can quantify their co- or anti-colocalization. The proposed method is a combination of the Maximum Entropy Method (MEM) and a Gaussian Copula, which we call the Maximum Entropy Copula (MEC). This new method can measure the spatial and nonlinear correlation of signals to determine the marker colocalization in fluorescence microscopy images.

RESULTS: The proposed method is compared with MEM for bivariate probability distributions. The new colocalization metric is validated on simulated and real data. The results show that MEC can determine co- and anti-colocalization even in high background settings. MEC can, therefore, be used as a robust tool for colocalization analysis.

AUTHORS/INSTITUTIONS: Z. Aminifarsani, statistics, Lorestan University, Khorramabad, IRAN (THE ISLAMIC REPUBLIC OF)
Through sequencing the 16S rRNA region, we could identify the composition of the microbiome that are associated with the health status of the individual. To detect microbiota which would contribute to each individual’s health condition, various statistical methods have been developed. Operational taxonomic units (OTUs), clustering sequences from microbiome samples by their sequence’s similarity, were able to measure the read counts. In biological classification point of view, OTUs could be expressed as various hierarchies called taxonomic rank. Using statistical methods with the taxon, it is possible to find the microbiota that affect human’s health condition. Recently, a novel method, HisCoM-microb, suggested using hierarchical structure of taxonomy information. HisCoM-microb used OTUs and taxonomic rank simultaneously. However, HisCoM-microb can only derive the effect on the phenotype at single taxonomy level. Moreover, analyzing only at a single taxonomy level does not imply which level of taxon are more strongly associated with the phenotype. To complement for these points, we propose a multi-level hierarchical component model to account for multiple taxonomic levels. In other words, we consider several taxonomic levels from OTUs in the model simultaneously. Through empirical studies, we compare the performance our method with HisCoM-microb and other existing statistical methods.
Abstract Body: One of the major goals in HIV clinical research is to identify homogeneous subgroups of patients for tailoring treatments within a personalized medicine perspective. From a statistical point of view, models accounting for unobservable heterogeneity in HIV patients are needed. While therapies have reduced cause-specific mortality in HIV patients, deaths due to other causes are increasing. In particular, when the focus is on personalized treatment, kidney disorders represent one major hurdle since HIV-positive individuals are at increased risk for chronic kidney disease (CKD). To prevent CKD, clinical guidelines recommend constant assessment of renal function in order to monitor nephrotoxic side effects drug-related. Glomerular Filtration Rate (eGFR) is commonly used as a renal function phenotype. We propose latent class mixed models (LCMMs) to model eGFR trajectories, which have the advantage, with respect to more standard approaches, of including different levels of heterogeneity in the population under study. This approach leads to the identification of different unobserved sub-populations with specific longitudinal response patterns. LCMMs consist of (i) a structural latent model, where the latent process is modelled as a function of time and other covariates without measurement errors; (ii) a measurement model, which links the latent process to the outcome of interest. In presence of heterogeneous population, a latent class-specific process is specified along with a multinomial logistic regression for the evaluation of latent class membership.

The key advantage of this modelling procedure is that, once clusters of subjects with similar eGFR dynamics are uncovered, relevant clinical information can be derived and applied in the clinical practice. LCMM performance evaluation is based on a sample of 1854 HIV infected patients. Different numbers of latent classes and BIC criterion has been considered for inference to select the optimal number of latent classes. To characterize patients in the uncovered latent classes we consider two different approaches: we use (i) the class membership model within LCMM framework and (ii) a Gaussian copula graphical model to explore the complex relationships among identified groups and the collected clinical characteristics. We compare findings from the two methods in identifying clinical features associated with patients assigned to different latent classes.

Authors/Institutions: C. Brombin, F. Cugnata, A. Castagna, C. Di Serio, Vita-Salute San Raffaele University, Milano, ITALY|L. Galli, A. Poli, A. Castagna, Clinic of Infectious Diseases, Istituto Scientifico San Raffaele, Milano, ITALY|C. Brombin, F. Cugnata, C. Di Serio, University Centre for Statistics in the Biomedical Sciences (CUSSB), Milano, ITALY
Spatial-temporal effects of the Integrated Health Systems Strengthening (IHSS) program on coverage of high impact child health interventions in Malawi.

ABSTRACT BODY:

Background: The Integrated Health Systems Strengthening (IHSS) program, which was implemented in six African countries: Ethiopia, Ghana, Malawi, Mali, Mozambique and Niger between 2007 and 2013, was aimed at increasing coverage indicators for high impact maternal and child health interventions. This study examined changes in the coverage of three high impact child interventions: measles vaccine, antimalarial treatment, and oral rehydration salts for diarrhea at the subnational level in Malawi before and after the IHSS program.

Methods: We analyzed data from 2004 and 2016 Demographic and Health Surveys, the former as a baseline and the later as end line data. We used multivariate Bayesian spatial models to estimate the three high impact child interventions jointly at the respective survey periods and to assess changes relative to the period between the two national surveys. We estimated the contribution of poverty and hospital delivery on changes in the three high impact child interventions.

Results: The odds were higher for each child's health intervention coverage after the IHSS implementation. The three high impact child health interventions varied substantially between districts and were associated with both district-level poverty and hospital delivery rates.

Conclusions: Increased coverage for the three high impact child health interventions, namely: measles vaccine, antimalarial treatment, and oral rehydration salts for diarrhea were observed in Malawi after the Integrated Health Systems Strengthening (IHSS) program. The changes varied across subnational levels, which might be driven by the local synergistic effect of major child health and policy strategies. The proposed statistical methodology has identified districts where changes in improved coverage indicators for child health interventions had taken place. This could help policymakers in targeting resources and programs to districts that are most in need or could be targeted to achieve improved coverage for high impact child health interventions.

AUTHORS/INSTITUTIONS: S.O. Manda, N. Haushona, Biostatistics Research Unit, South African Medical Research Council, Pretoria, Gauteng, SOUTH AFRICA|S.O. Manda, Department of Statistics, University of Pretoria, Pretoria, SOUTH AFRICA|
Abstract Body: Previous research in the USA has established a link between mental health problems in youth and tobacco use. However, it is unknown if mental health problems lead to earlier ages of e-cigarette initiation in youth. The Population Assessment of Tobacco and Health (PATH) study used the Global Appraisal of Individual Needs-Short Screener to measure past-year symptoms of internalizing and externalizing disorders. Internalizing disorder was characterized by measuring sleeping restlessly, anxiety, depression, and distress. Externalizing disorder was characterized by measuring lying, hard time paying attention or listening to instructions at school, bullied or threatened other people, started physical fights, feeling restless and giving answers before other people finish asking. Response options on all questions were never, more than 12 months ago, 2-12 months ago, and within the past month. The sum of the number of symptoms in the past year was estimated and then collapsed participants into low (0 score), moderate (1-2), and high (3+ symptoms) probability of diagnosis for each scale. Survival analyses for interval-censored data from 2013-2017 of PATH youth (n=16,143; N=29,349,766) were used to determine the age of initiation of e-cigarettes and the association of each scale with the age of e-cigarette initiation. Hazard ratios and 95% confidence intervals were estimated, adjusting for the interaction of sex and race. For the internalizing scale, among youth who scored low or moderate, 14% initiated e-cigarettes by age 16, and among youth who scored high, 14% initiated e-cigarettes by age 15. For the externalizing scale, among youth who scored high, 25% initiated e-cigarettes by 18 years, while youth who scored moderate or low, 25% initiated e-cigarettes by 17 years. The risk of initiating e-cigarettes at earlier ages is 1.94 (95%CI=1.93-1.95) times higher in participants who scored high on the internalizing scale versus those who score low on the internalizing scale. The risk of initiating e-cigarettes at earlier ages is 2.29 (95%CI=2.28-2.30) times higher in participants who scored high on the externalizing scale versus those who score low on the externalizing scale. This study provides evidence that internalizing and externalizing disorders predispose youth to earlier ages of e-cigarette initiation.

Authors/Institutions: M. Bluestein, C. Perry, M. Harrell, Michael & Susan Dell Center for Healthy Living, UTHealth, Austin, Texas, UNITED STATES|B. Chen, A. Perez, Biostatistics and Data Science, The University of Texas Health Science Center at Houston, Austin, Texas, UNITED STATES|
Storm surge is a rise of the sea level as a result of wind associated with atmospheric pressure changes or storms. Storm surge combined with inland rainfall and high tide results in large impact on coastal areas and natural and human habitants. Understanding wind force measured from scattered buoys is critical to statistical modeling of storm surge. One challenge of using observed wind speed data from multiple stations is handling dependency among buoys. This study aims to model spatio-temporal dependency in multivariate wind speed data, which possess time-varying volatility. Our study introduces multivariate GARCH models such as dynamic conditional correlation, GARCH-copula, and generalized orthogonal factor GARCH models to model volatile multivariate wind speed. The GARCH type volatility model remains unflagged in spatio-temporal modeling of environmetrics. The DCC models depend on the choice of univariate GARCH, dynamic multivariate correlation matrix structure and distribution of error term. The GARCH-copula and GOF-GARCH models additionally require elliptical distributions and orthogonal factor decomposition, respectively. The models are applied to the wind speed data measured at multiple buoys along the Carolina coastal line which include hurricanes. After information criterion based model selection, the results demonstrate unique time-varying correlations among multiple buoys. This dynamic correlation study is expected to help understand dynamic relationships in wind speed data under various weather conditions and estimate spatial interpolation of the dynamic correlation.
The Nun Study, a longitudinal study to examine risk factors for the progression of dementia, consists of subjects who were already diagnosed with dementia (i.e., prevalent cohort) and those who do not have dementia (i.e., incident cohort) at study enrollment. When assessing the risk factors’ effects on the survival time from dementia diagnosis until death, utilizing data from both cohorts supports more efficient statistical inference because the two cohorts provide valuable complementary information. A major challenge in analyzing the combined cohort data is that the prevalent cases are not representative of the target population. Moreover, the dates of dementia diagnosis are not ascertained for the prevalent cohort in the Nun Study. Hence, the survival time for the prevalent cohort is only partially observed from study enrollment until death or censoring, with the time from dementia diagnosis to study enrollment missing. In this paper, we propose an efficient estimation method that uses both incident and prevalent cohorts under the proportional mean residual life model. By assuming proportionality of the mean residual life time with covariates in the incident cohort, we can utilize the natural relationship between the mean residual life function and the hazard function of the survival time measured from enrollment until death for the prevalent cohort. We evaluate the efficiency gain from using the combined cohort data through simulations and demonstrate that the proposed method is valid and efficient.

C. Lee, Biostatistics and Epidemiology, University of Massachusetts, Amherst, Massachusetts, UNITED STATES
J. Ning, Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, UNITED STATES
Y. Shen, Biostatistics, UT MD Anderson Cancer Center, Houston, Texas, UNITED STATES
R.J. kryscio, Biostatistics, University of Kentucky, Lexington, Kentucky, UNITED STATES
Abstract: Economists and policy analysts regularly compare treatments on the basis of both cost and effectiveness in order to inform resource allocation decisions and health policy. Several statistical challenges impede our ability to compare treatments on the basis of cost-effectiveness. First, these analyses are often conducted using observational data and, therefore, necessitate confounding control. Second, costs accumulated under a treatment and the treatment's efficacy (often defined as a survival outcome) tend to be censored. Third, cost and effectiveness are necessarily correlated - requiring joint modeling of a very complex, irregular distribution plagued with skewness, multimodality, and structural zeros. Our work has several contributions. We formally identify causal cost-effectiveness contrasts under modified ignorability and consistency assumptions. We then develop a flexible, nonparametric Bayesian model for the joint distribution that accommodates the complexity of these outcomes. This approach models the joint outcome via an adaptive mixture of simpler, parametric regression models. It is adaptive in the sense that the number of mixture components need not be pre-specified. Instead, as many components are introduced as the complexity of the distribution warrants. We incorporate this model into a Bayesian g-formula to compute the identified causal cost-effectiveness contrasts. Finally, we show how the model-based clustering of observations into mixture components can be used to automatically detect subgroups of the target population with heterogeneous cost-effectiveness profiles. This provides an advancement over standard methods which compute cost-effectiveness contrasts for pre-specified (not adaptively discovered) subgroups, then check to see which subgroup differences are meaningful. We outline an MCMC scheme for posterior sampling, assess frequentist properties our estimator under a variety of scenarios via simulation, and apply use our model to analyze the cost-effectiveness of various treatments among patients with endometrial cancer.
Abstract Body: An Isolation-with-Migration (IM) model explains the genetic divergence of two populations split away from their common ancestral population. Under an IM model, a genealogy consists of two kinds of evolutionary paths of genetic data: vertical inheritance paths through generations and migration paths between populations. The computational complexity of IM model inference is one of the major limitations to analyze genomic data. We propose a fast maximum likelihood approach to estimate IM models from genomic data. The first step analyzes genomic data and maximizes the likelihood of a coalescent tree that contains vertical paths of genealogy. The second step analyzes the estimated trees and finds the parameter values of an IM model, which maximizes the distribution of the coalescent trees after taking account of possible migration events. We evaluate the performance of the new method by simulated and real data analyses.
Abstract Body: Meta-analysis is a powerful tool for drug safety assessment by synthesizing findings from independent clinical trials. However, a common challenge is that a large number of published clinical studies may not report rare adverse events. If fewer events are observed than a pre-specified cutoff, these events may not be reported in the publication. To derive exact inference and robust estimates for the missing not at random data, we propose a Bayesian multilevel regression model in conditional coarsened data framework to accommodate censored sparse binomial event data. Given a stochastic data-coarsening mechanism, a sensitivity analysis is also suggested to assess whether it is appropriate to ignore the stochastic nature of the coarsening. The proposed approach is illustrated using data from a recent meta-analysis of 125 clinical trials involving PD-1/PD-L1 inhibitors with respect to their toxicity profiles. We demonstrate that if the censored information is ignored, the incidence rate of adverse event is overestimated; this bias could have significant impact on immunotherapy drug adoption and public health policy.

Authors/Institutions: S. Zhou, Public Health Sciences, Pennsylvania State University, Hershey, Pennsylvania, UNITED STATES|X. Qi, C. Peterson, Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, UNITED STATES|
ABSTRACT BODY:

**Abstract Body:** Rapid developments in spatial streaming data technologies are continuing to generate increased interest in monitoring human activity cycles. Among prevalent non-invasive practices for measuring gross motor activity are sensor units worn on the wrist (actigraphy). An actigraph unit continually records the activity level of an individual, producing a very large amount of data at a high-resolution that can be immediately downloaded and analyzed. While this kind of "BIG DATA" includes both spatial and temporal information, the variation in such data seems to be more appropriately modeled by considering stochastic evolution through time while accounting for spatial information separately. Our current work develops a comprehensive Bayesian hierarchical modeling and inferential framework for actigraphy data reckoning with the massive sizes of such databases while attempting to offer full inference. Building upon recent developments in modeling massive spatial-temporal data, we construct Nearest Neighbor Gaussian Processes (NNGPs) for actigraphy data to compute at massive temporal scales while accounting for spatial information using spline regression. More specifically, we construct a temporal NNGP and we focus on the optimized implementation of the collapsed algorithm in this specific context. This approach helps us scale up the models while also offering full inference. We test and validate our methods on simulated data and subsequently predict missing or future activity level for the subjects included in a data set coming from a study conducted by environmental health scientists at the Fielding School of Public Health in the University of California Los Angeles. The data comprises activities of a cohort of 123 individuals who have been monitored over a period of two weeks.

**AUTHORS/INSTITUTIONS:** M. Mingione, P. Alaimo Di Loro, Statistics, University of Rome "La Sapienza", Rome, Lazio, ITALY | M. Mingione, Institute of Applied Calculus "M. Picone", National Council of Research, Rome, Lazio, ITALY | M. Mingione, P. Alaimo Di Loro, S. Banerjee, J. Lipsitt, Biostatistics, University of California, Los Angeles, Los Angeles, California, UNITED STATES
Title: Making it Easy to Produce Climatic Visualizations Using ggplot2 in R-Instat.

Abstract Body: Ggplot2 is an excellent and very powerful data visualisation package in an increasingly popular R. It is an implementation of the Grammar of Graphics by Wilkinson. Despite its brilliant capabilities, a lot of people including experienced R programmers find it difficult to use. R, being a command driven language, it has a very steep learning curve. Ggplot2 being a package in R, many people who need the visualisation capabilities provided by ggplot2 for their work shy away from using it because it requires some programming understanding to produce the visualisations. As there are many front ends to R, meant to make it easy for people to use R, none of them offers the climatic capabilities that meteorology data people need to easily and quickly produce graphs. To solve these challenges, we have developed R-Instat, which is a front end to R, that is menu driven, free and open source. It is intuitive and very easy to use too. We have developed user interfaces that are easy to use to produce useful climatic graphs from historical climatic data. These graphs include start of rains, length of season, water balance and rainfall totals which are very useful in Agrometeorology. We hope that through interactions with R-Instat, then users may use it as a learning tool to learn R. This paper will describe the ideas behind development of R-Instat, how we have enabled easy use of ggplot2 and the impact it has had in improving production of climatic products useful for agriculture across a number of African countries.

Authors/Institutions: M.F. Njiru, School of Computing and Informatics, Maseno University, Maseno, Kenya | M.F. Njiru, African Maths Initiative, Kisumu, Nyanza, Kenya |
Abstract Body: Epilepsy affects millions of individuals worldwide and uncontrolled seizures are a major burden on resources. Whilst seizures are effectively controlled by anti-epileptic drugs (AEDs) in a majority (~70%) of individuals, there is significant heterogeneity in AED response. Many fail to achieve remission and others develop adverse drug reactions (ADRs) necessitating treatment withdrawal. Genome-wide association studies (GWAS) have been largely unsuccessful so far in identifying genetic determinants of AED response.

Whole-exome sequencing (WES) is enriched for protein-coding variants affecting molecular function with direct biological interpretation, includes low-frequency and rare variants not typically analysed in GWAS. It is therefore anticipated that a WES approach will be more successful in identifying variants explaining variability in AED response. Statistical methods and software to test association with GWAS/WES variants are largely aimed at binary and quantitative traits, being the most common outcomes in disease genetics. However, when studying genetics of treatment response (pharmacogenetics), outcomes such as time to ADRs and treatment remission are often most important and there exists a lack of analysis tools aimed at such outcomes in both GWAS and WES settings. To address this analytical bottleneck, our group is focussed on developing methodologies and software capable of analysing GWAS and WES datasets with survival outcomes.

In this study, we analyse WES and GWAS data with time to event outcomes representing variable AED response from over 1300 individuals with epilepsy, applying software packages SurvivalGWAS_SV and rareSurvival, recently developed by our group specifically for survival endpoints. Results are compared to those obtained from applying more standard methods and software. The methodology and results for more complex survival outcomes that include competing risks to allow for different endpoints, will also be discussed.

Authors/Institutions: R.G. Shankar, A. Morris, A. Jorgensen, Biostatistics, University of Liverpool, Liverpool, UNITED KINGDOM|T. Marson, Molecular and Clinical Pharmacology, University of Liverpool, Liverpool, UNITED KINGDOM|G. Sills, School of Life Sciences, University of Glasgow, Scotland, UNITED KINGDOM|A. Morris, School of Biological Sciences, University of Manchester, Manchester, UNITED KINGDOM
On-farm experimentation (OFE) has been proposed as a process to enhance adoption of digital technologies in agriculture, develop farmer competence and improve farm profitability. OFE is a farmer-centric process where farmers work with consultants or researchers to design and implement experiments to test management practices. Farmers use their own machinery to conduct large field-scale experiments. Data are usually measured by yield monitors that are standard in most modern harvesting machinery. However, data recorded by other types of sensors mounted on machinery, satellites or drones may also be used.

Agricultural experiments have historically been performed by researchers using small plot sizes and principles of experimental design to so that effects of treatments can be analysed while minimising or accounting for environmental effects. In contrast, analysis of data from large on-farm experiments is challenging because of large spatial variation in yield that: (1) is not due to the treatment being tested; (2) is influenced by spatial auto-correlation; and (3) is often much larger than the variation due to treatment effects. In addition, the relationship between treatment and yield may also vary spatially.

We investigate the use of Bayesian latent Gaussian modelling (BLGM) for analysis of data from large on-farm experiments. Latent Gaussian models combine random Gaussian effects (that may be fixed, structured or unstructured) into a linear predictor. We consider the case where the structured random effects include spatial dependency modelled as a Gaussian field with Matérn covariance. Bayesian inference is performed using integrated nested Laplace approximation. Examination of the posterior densities for the Matérn parameters shows they can be accurately estimated from the data using this approach.

We also consider the issue of model selection for BLGM; in particular determining whether treatment effects can be modelled globally or whether a spatially varying coefficient (SVC) model is required. We compare three approaches to model selection: (1) ad-hoc bandwidth based on experimental design; (2) k-fold cross validation; and (3) spatial k-fold cross-validation.
Abstract Body: As smart devices and digitalization in every corner of life become commonplace even for ordinary citizens, statisticians and data scientists need to address the growing need of how to handle the massive amount of streaming data and turn them into information and then scientific knowledge as efficiently and quickly as possible. Streaming healthcare data is a good example of this as well. In this presentation, we plan to present the National HealthCare data, how we approach evidence-based health policy development, and how to structure real-time feedback systems to increase the visibility of accumulation data to the treating physician and thus help improve the health outcomes in a timely manner.

AUTHORS/INSTITUTIONS: M. Kocak, Regenerative and Restorative Medicine Research, Istanbul Medipol University, Beykoz, Istanbul, TURKEY
The detection of genomic region involved in local adaptation is, arguably, one of the main challenges of modern population genetic.

Many statistical procedures have been developed to detect them. All are computationally slow. This hamper their application to high-throughput SNP genotypic datasets. Indeed, on such datasets their runtime is in hours if not days.

Most procedures rely on a divergence score based on population allelic frequencies measured at each marker (e.g. Fst). This score is then analyzed either
(i) through multivariate procedures that perform a joint analysis of the score but do not account for the spatial arrangement of the markers (e.g. Bayesian Factor models) or
(ii) through genome scan procedures that account for the spatial distribution of the signal (e.g. BayeScan).

We follow this last line of research and propose a new genome scan approach based on the robust Fst estimator recently proposed by Bhatia. We demonstrate that the problem of segmenting the genome into Fst-homogeneous regions can be recast as a weighted changepoints detection problem. When comparing two populations we propose to use the quasi-linear, yet exact fpop-algorithm, to recover the penalized maximum likelihood estimator. It runs in a matter of minutes for a 1KG dataset. This drastically reduces the computational burden while retaining strong statistical guarantees. We illustrate our approach on the 1KG dataset.

AUTHORS/INSTITUTIONS: T. Mary-Huard, G. Rigaill, INRA, Paris, FRANCE
Abstract Body: In epidemiology and medicine, the relative risk is an important quantity of interest. Log-binomial regression is a popular modelling strategy to model the relationship between a binary outcome and a set of covariates; in such models, the coefficients of the covariates represent the adjusted log relative risks, and modified Poisson regression can be used to obtain their empirical estimates. However, missing data are prevalent in many medical and epidemiological datasets for a variety of reasons. For example, patients participating in longitudinal studies may drop out before its conclusion, or subjects may omit responses in questionnaires. Missing data can result in reduced statistical power, biased estimation of parameters, and weakened generalizability of results. Sullivan et al. (BMC Medical Research Methodology, 2017; 17: 134) considered a setting with a binary outcome and continuous covariates subject to missingness characterized by a log-binomial model where the true adjusted relative risks were known. They handled the missing data with multiple imputation by fully conditional specification (FCS), using logistic regression to impute for binary variables and linear regression to impute for continuous covariates. They found that using these popular but mis-specified imputation models to impute missing values followed by estimating the adjusted relative risks with a correctly specified log-binomial model often resulted in downward bias and incorrect coverage. We propose a novel FCS multiple imputation procedure to impute for binary variables and continuous covariates. First, the proposed method imputes binary variables with a procedure that combines discretization with local logistic regression adjustment. Second, the proposed method imputes continuous covariates using a procedure that uses discretization followed by imputation under a conditional distribution that is “compatible” (Bartlett et al., 2015, Statistical Methods in Medical Research; 24(4): 462-487) with a logistic regression model for the outcome within each sub-interval. Using extensive simulations, we demonstrate that our novel FCS multiple imputation procedure achieves unbiased estimation with nominal coverage in a wide variety of scenarios when estimating adjusted relative risks using a correctly specified log-binomial regression model.
Hypertrophic cardiomyopathy (HCM) is a cardiovascular disease with a prevalence of 1 in 500. It is the most frequent cause of sudden death in young people and can lead to other heart diseases. HCM is primarily caused by mutations in several genes, but then modulated by other factors, such as epigenetics and gene expression. Our motivating data consists of 15k transcription and 30k methylation levels, respectively, in 13 HCM cases and 10 controls. Standard univariate tests suffer from a lack of power due to the multiple testing issues. Penalized regression methods consider each omics dataset separately, without exploiting the relations between both datasets. A supervised integrated analysis involving all omics datasets provides a more holistic view of the pathophysiology behind HCM and more insight in inter-omics relations.

Our aim is to develop a supervised framework to integrate transcriptomic and epigenomic variation and simultaneously relate it to HCM. Methodological challenges are high dimensionality (p >> N) and the presence of technology-specific heterogeneity. Methods that have been proposed for two omics datasets and a continuous outcome are Collaborative Regression (CoRe), IPF-LASSO, and LRM. They formulate a least squares objective function and apply penalized regression to find coefficients for each feature that can predict the outcome. The LRM approach additionally utilizes correlations across the omics data by including latent components shared between datasets in the regression. In these methods, the technology-specific heterogeneity is not properly addressed and overfitting can occur when applying least squares methods to high dimensional data.

We propose supervised PO2PLS (sPO2PLS), a supervised probabilistic model for the relation between two datasets and an outcome. The model includes joint and specific latent variables that capture correlations between omics data and address technology-specific heterogeneity, respectively. The association with the outcome is modeled via these latent variables. The components and coefficients are estimated with maximum likelihood. Inference is facilitated within the sPO2PLS model.

An extensive simulation study will be conducted to investigate the performance of sPO2PLS compared to alternative methods. We apply sPO2PLS to the omics data to detect the most relevant features for separating HCM cases from controls.

**AUTHORS/INSTITUTIONS:** G. Naylor-Leyland, H. Uh, J. Houwing-Duistermaat, S. El Bouhaddani, Biostatistics and Research Support, University Medical Center Utrecht (UMCU), Utrecht, NETHERLANDS|J. Houwing-Duistermaat, Department of Statistics, University of Leeds, Leeds, UNITED KINGDOM|
Single-cell RNA sequencing technology provides an opportunity to study gene expression at single cell resolution. However, prevalent dropout events in the data cause high sparsity and noise level that obscure downstream data analysis. We propose a gene-graph-based imputation method, G2S3, that imputes for dropouts by borrowing information from adjacent genes in a sparse gene graph learned from the data. We applied G2S3 and other imputation methods to several datasets to assess and compare their performances. Results showed that G2S3 is superior in recovering the true gene expression level, identifying true cell subtypes and stages, and improving differential expression analyses.
The use of Artificial Intelligence (AI), machine learning and deep learning have gained a lot of attention and become increasingly popular in healthcare research. Historically, machine learning and theory had strong connections to statistics; however, the current deep learning context is mostly in computer science perspectives and lacks statistical perspectives. In this talk, we address this research gap and discuss how to teach deep learning to the next generation of biostatisticians. We first describe some backgrounds and how to get motivated. We discuss different terminologies in computer science and statistics, and how deep learning procedures work without getting into mathematics. In response to a question regarding what to teach, we address organizing deep learning contents and focus on the statistician’s view; form basic statistical understandings of the neural networks to the latest hot topics on uncertainty quantifications for prediction of deep learning, which has been studied in the Bayesian frameworks. Further, we discuss how to choose computational environments and help develop programming skills for the students. We also discuss how to develop homework incorporating the idea of experimental design. Finally, we discuss how to expose students to the domain knowledge and help to build multi-discipline collaborations. We finish this talk by illustrating real-world biomedical applications and envisioning future opportunities in cancer research and AI in healthcare.

This talk is based upon a working paper in progress from the deep learning working group at the Statistical and Applied Mathematical Sciences Institute (SAMSI), USA.

Hyunsoon Cho’s work was partially supported by the National Research Foundation of Korea grant (No. NRF-2019R1H1A1079981).

**AUTHORS/INSTITUTIONS:** H. Cho, Cancer Control and Population Health, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang, KOREA (THE REPUBLIC OF)|H. Cho, J. Babu, D. Banks, Statistical and Applied Mathematical Sciences Institute, Durham, North Carolina, UNITED STATES|J. Babu, Statistics, Pennsylvania State University, University Park, Pennsylvania, UNITED STATES|D. Han, University of Texas at San Antonio, San Antonio, Texas, UNITED STATES|D. Banks, Statistical Science, Duke University, Durham, North Carolina, UNITED STATES|
TITLE: A novel statistical method for integrating multi-omics data

Abstract Body: Alteration in regulatory activities of a gene may be attributed to various genetic and epigenetic factors, like, mutation in DNA, methylation at various CpG sites etc. Traditional methods of analysing omics data separately would incur loss of essential information that might be beneficent in differentiating between the genetic code of a tumor and a normal sample or for case-control data. Recently data integration through joint analysis of various omics data is focused to obtain robust inferences. But it is imperative to understand and include the available downstream information and use more practical assumptions in the model development, in order to capture the diversity in the genomic profiles of patients. We have developed a new statistical method that integrates information from genotype data and gene expression data along with other covariates under one framework. Thus our method includes a biological insight from multi-omics data to assess its effect on the gene expression. We also derive the asymptotic distribution of our test statistic to calculate p-values fast. We perform extensive simulations to analyse the performance of our method mainly in terms of size and power of the test. The type I error rate, calculated on the basis of the distribution of the test statistic, is controlled at 5% level of significance. Our proposed method is powerful and consistent as the power of the test increases with the increase in sample size. Our method is robust as it shows consistent results for different genetic models that we consider in our simulation study. We also apply this method to a real dataset and found few novel SNPs. Based on the results obtained from simulated and real data, our method looks very promising in differentiating gene expression profiles of patients through integrated analysis of multi-omics data.

AUTHORS/INSTITUTIONS: I. Mukhopadhyay, Human Genetics Unit, Indian Statistical Institute, Kolkata, WEST BENGAL, INDIA.
Metabolic syndrome is considered a cluster of health outcomes that results from a combination of increased blood pressure, high glucose concentration in the plasma, excess abdominal fat and high cholesterol levels. Metabolic syndrome is a latent construct which is not directly observable. However, it is suspected or diagnosed based on the presence of multiple health outcomes or indicators. In studying the causal pathway to the development of metabolic syndrome, latent variable models such as the multiple indicators, multiple causes models are applied. In such settings, researchers can better understand the potential pathway that exists between exposure to a high-fat diet, physical activity, and metabolic syndrome development. The Multiple Indicators, Multiple Causes Measurement Error (MIMiC ME) model is used to study the impact of an underlying latent construct on its multiple outcomes where the potential causes of the underlying latent construct are measured with error. In this manuscript, we extend the MIMiC ME model to allow a functional covariate measured with error and we study the impact of metabolic syndrome on its multiple indicators as well as the role of exposure to a high-fat diet and total daily energy expenditure on the development of the syndrome. We adopt a Bayesian Semi-parametric approach to estimate the model parameters. Simulation studies are also performed to assess sensitivity to violations of our model assumptions. The newly defined model is applied to study the impacts of daily energy expenditure and exposure to a high-fat diet among the ZDF rats, a genetically obese animal model.
Abstract Body: Practical considerations lead to the use of unit of analysis within subjects, e.g., bleeding episodes or treatment-related adverse events, in rare disease settings. This is coupled by data augmentation techniques such as extrapolation to enlarge the subject base. In general, one can think about extrapolation of data as extending information and conclusions from one estimand [Akacha, M., et. al. (2017)] to another estimand. This approach induces hierarchical clustered data with varying cluster sizes.

Extrapolation of clinical trial data is being accepted increasingly by regulatory agencies as a means of generating data in diverse situations during drug development process. Under certain circumstances, data can be extrapolated to a different population, a different but related indication, and different but similar product. We consider here the problem of estimation (point and interval) using a mixed models approach under an extrapolation.

We propose construction of estimators (point and interval) from two methods – frequentist (using weighting schemes for the clusters, e.g., equally weighted and with weights proportional to cluster size) and Bayesian [using power prior model (Gamalo-Siebers et al., 2017)]. We then evaluate the performance of this approach using simulated data under varying scenarios.

In conclusion, we see that the approach is a useful means for improving statistical inference in rare disease settings and thus aids not only signal detection but risk-benefit evaluation as well.

Authors/Institutions: D.I. Bonzo, E. Wang, J. Prescod, Biometry, LFB, Nashua, New Hampshire, UNITED STATES
Estimating the difference in restricted mean survival time accounting for trial effect in individual-patient-data meta-analyses

Abstract Body
The difference in restricted mean survival times (RMSTs) between two treatment groups is a useful tool to provide information on the average causal treatment effect in a randomized clinical trial. This method is particularly appealing since it does not require a proportional hazards assumption, unlike the hazard ratio based on the Cox model. The RMST can be obtained by integrating under the survival curve up to a predetermined horizon. A particular concern in individual patient data meta-analyses (IPD-MA) is to properly account for the trial effect in the estimation of the difference in RMSTs.

Our objective is to estimate the difference in RMSTs in an IPD-MA using various propositions where the RMST is adjusted on covariates and includes a random trial effect.

Method(s): In our 1st proposition, we use the Breslow estimator as proposed by Zucker (JASA 1998) to have an estimation of the baseline hazard adjusted on the trials and on other covariates, and calculate the area under the curve. A variance estimator is obtained with an added term in the expression (Chen and Tsiatis Biometrics 2001) to remove the conditioning on the trial. The 2nd proposition is based on a general linear mixed model which includes a random trial effect. A Poisson regression allows the estimation of the survival baseline hazard. Finally, the 3rd proposition is based on a Breslow estimator which takes into account the random trial effect (Gorfine et al. Biometrika 2006). An IPD-MA from the GASTRIC (JAMA 2010) is used for illustration.

Results: A simulation study has been set up generating Weibull data with random trial and treatment-by-trial effects. A true “RMST” is calculated through a double integration of the parametric survival function; and the bias and variance of our estimation methods are evaluated with respect to this value. We also plan to make simulations under a non-proportional hazard hypothesis. The IPD-MA included 3288 patients with resectable gastric cancer from 14 randomized trials of adjuvant chemotherapy versus surgery alone. At the 10y [20y] horizon, the estimated difference in RMSTs is 141[354] days (SE 45 [104] days; p=0.002 [p<0.001]) between the chemotherapy and surgery alone groups.

Conclusions: We proposed estimation methods for the difference of RMSTs between two treatment groups in the IPD meta-analyses context.

Abstract Body: The increase in the use of sensor-enabled agriculture equipment and field image data has enabled farmers to perform their own experiments using implementable but informative designs and analyses. Sensor technology is changing both design and analyses, making geostatistical approaches that exploit the spatial dependence and explicit spatial information attractive. It is allowing farmers to understand and manage covariates (e.g. pH, topography, … ) more easily and consider those in the design. Can these factors reduce the need for randomisation? The type of information of interest to farmers is more about the spatial pattern of treatment effects not so much on average field-level differences. Can on-farm experimentation deliver pragmatic, but statistically rigorous information that will encourage farmers to perform their own experiments?

Authors/Institutions: R. Darnell, B.L. Henderson, Data61, CSIRO, Brisbane, Queensland, AUSTRALIA