

## Found 356 Records

**CONTROL ID:** 3276619

**TITLE:** Some New Nonlinear Hyperbolic Growth Equations For Modeling Tree Growths

**ABSTRACT BODY:**

**Abstract Body:** Studies have shown that majority of the growth models emanated from the Malthusian Growth Equation (MGE), which is limited to growing without bounds. This study was designed to develop alternative growth models flexible to enhance internal prediction of biological processes based on hyperbolic sine function with bound. The intrinsic rate of increase in the MGE and its variants were modified by considering a growth equation, which produces flexible asymmetric curves through nonlinear ordinary differential equations of the form;  $dH/dt=H[r+\theta/\sqrt{(1+t^2)}]$ . The developed hyperbolic growth models captured boundedness in Malthusian Growth Equation, improved general fitness and robustness over exponential, monomolecular, Gompertz, Richards and von Bertalanffy growth models. The developed hyperbolic growth models improved general fitness over exponential, monomolecular, Gompertz, Richards and von Bertalanffy growth models in predicting tree heights, diameter at breast heights using Tree Age. Also, published datasets used demonstrated effectiveness of the developed models in the prediction of the response variable.

**AUTHORS/INSTITUTIONS:** O.S. Oyamakin, Statistics, University of Ibadan, Nigeria, Ibadan, Oyo, NIGERIA|

**CONTROL ID:** 3276680

**TITLE:** DISCLOSING FACTORS BEHIND HIV/AIDS PREVALENCE: CROSS SECTIONAL STUDY ON EAST AND SOUTH SUB-SAHARA AFRICAN COUNTRIES

**ABSTRACT BODY:**

**Abstract Body:** Author: Ombeni Eliapenda Kaluse

A Graduate of Bachelor of Official Statistics, Eastern Africa Statistical Training Centre (EASTC)-Tanzania

**Keywords**

HIV/AIDS prevalence rate, Unemployment, Population growth, Gross Domestic Product (GDP) and Net enrolment in primary schools. .

**ABSTRACT**

Acquired Immune Deficiency Syndrome (AIDS) is a disease caused by HIV virus that affects the human immune system tremendously eventually leading to death. HIV is considered as one of the fatal cause of death in the present times. The global scenario of AIDS is alarming and number of infected patients is regularly increasing. (ISSN 2155-6113, Journal of AIDS & Clinical Research, 2014).

By 1987, the epidemic began gradually to move southern parts of Africa. Some of the most explosive epidemics have been seen in Southern Africa. South Africa has the largest number of people living with HIV/AIDS in the world of about 5 million. Botswana and Swaziland have the highest prevalence levels, 38% and 33% respectively. West Africa has been relatively less affected by HIV infection than other regions of sub-Saharan Africa where by Botswana, Swaziland, Tanzania, Uganda and Kenya are located. Reported by UNAIDS and WHO. (Research, 2014 UN/POP/MORT/2003/2, 5 September 2003).

The causative factors of the prevalence mostly reported to be the social economic factors of Unemployment, population growth, GDP growth and Enrolment of young people to Education have shown impacts to the vulnerability of HIV/AIDS prevalence across the East and southern Africa sub-Saharan countries.

The panel data analysis results from the Random Effect (RE) model, have proved the significance of the impact of the mentioned factors to HIV/AIDS prevalence rate whereby R-Square is 0.8938 with the Corr (Uit, Xit) = 0. The enrolment to primary school, unemployment and the GDP growth have shown positive correlation with the HIV/AIDS prevalence while Population growth has indicated negative correlation. This means that rapid population increase in the sub-Saharan countries can hide the HIV/AIDS prevalence rate over population but is not to say that HIV/AIDS is eradicated, it could have prevailed in a way we do not consider.

**AUTHORS/INSTITUTIONS:** O.E. Kaluse, Official statistics, Eastern Africa Statistical Training Centre , Dar-Es-Salaam, Changanyikeni, TANZANIA, UNITED REPUBLIC OF]

**CONTROL ID:** 3276828

**TITLE:** COMPARISON OF COX PROPORTIONAL HAZARD MODEL AND ACCELERATED FAILURE TIME MODEL WITH APPLICATION TO DATA ON TUBERCULOSIS/HIV PATIENTS IN NIGERIA

**ABSTRACT BODY:**

**Abstract Body:** In this research, we adopted various techniques such as the survival function curve, the Cox proportional hazards (PH) model, omnibus test, survival function of the mean covariates, log minus log, the Accelerated failure time (AFT) model, the AFT model plot, the Log-likelihood test and Akaike Information Criterion (AIC). So also we used the log-rank test for comparing all survival variables curves. These techniques were used for analyzing survival data on Tuberculosis/HIV co-infected patients in Nigeria. We apply the methods to a cohort of these patients managed in tertiary Directly Observed Treatment Short Course (DOTS) centre, Nigerian Institute of Medical Research (NIMR) for the period of six months, where we compare the effect of the accelerated failure time model with Cox proportional hazard model in determining the time to sputum conversion in TB patients who are co-infected with HIV. The research established that AFT model provides a better description of the dataset as compared with Cox PH model because it allows prediction of Hazard function, survival functions as well as time ratio. Moreover, PH model does not fit appropriately when compared with AFT model; thereby provide less appropriate description of survival data. The result revealed that the Weibull AFT model provided a better fit to the studied data than the Cox proportional hazards model. Hence, it is better for researchers of TB/HIV co-infection to consider AFT model even if the proportionality assumption of the Cox model is satisfied.

**AUTHORS/INSTITUTIONS:** O.O. OGUNGBOLA, A.A. Akomolafe, DEPARTMENT OF STATISTICS, FEDERAL UNIVERSITY OF TECHNOLOGY AKURE, ONDO STATE, Akure, Ondo, NIGERIA|A.Z. Musa, Monitoring and Evaluation, National Institute of Medical Research, Yaba, Lagos State, Yaba, Lagos, NIGERIA|

CONTROL ID: 3277290

**TITLE:**

Comparison of Two One-Sided Tests with Power Approach for Bioequivalence in Phase 1 clinical Trials

**ABSTRACT BODY:**

**Abstract Body:** The bioequivalence problem has always received considerable attention in the statistical literature. The statistical test of the hypothesis of no difference between the average bioavailabilities of two drug formulations, usually supplemented by an assessment of what the power of the statistical test would have been if the true averages had been inequivalent, continues to be used in the statistical analysis of bioequivalence (BE) studies. The discussion intends to cover this Power Approach, proposed by Hauck and Anderson (1987), which in practice usually consists of testing the hypothesis of no difference at  $\alpha$ -level 0.05 and requiring an estimated power of 0.80. This is compared to the Two One-Sided Tests (TOST) procedure, which leads to the same conclusion as the approach proposed by Westlake (1985) based on the usual (shortest)  $1-2\alpha$  confidence interval for the true mean difference. With appropriate choice of the nominal significance level for the one-sided tests, the TOST procedure always has uniformly superior testing properties to the power approach.

In power approach, the null hypothesis of no difference between the two treatment means, as tested by the variability due to treatment differences using an F-test from the analysis of variance of a two-treatment formulation study, is the wrong statistical hypothesis for assessing the evidence in favor of a conclusion of equivalence. This is because equivalence may not always represent equality. On the other hand, for TOST the statistical hypotheses are referred to as the "interval hypotheses", where the null hypothesis represents no equivalence against the alternative which considers equivalence.

Simulation results show the superiority of TOST in terms of the 90% confidence interval (CI) for BE as compared to Power approach. The TOST approach obtained BE based on shortest CI at  $\alpha = 0.05$ , whereas the power approach resulted in much wider CI. Comparative simulation analysis concludes that for showing BE, TOST can be consider as a better approach.

Comparison of the methods have been further illustrated using a real-life Phase 1 single-center study whose design is open-label, single dose, randomized, 4-period, 2-sequence fully replicated crossover study to assess BE of two treatments administered orally to healthy volunteers. While testing BE using both the procedures, similar conclusion has been achieved.

**AUTHORS/INSTITUTIONS:** A. Poddar, Biostatistics, IQVIA, Bangalore, INDIA|A. Mukherjee, Biostatistics, PPD International, Bangalore, INDIA|

**CONTROL ID:** 3278923

**TITLE:** The behavior between income, work, education and disability in a comparative study using regression techniques, profile graphs, municipal data, variable and model selection

**ABSTRACT BODY:**

**Abstract Body:** Abstract: Education contributes to the formation of human capital, being a determinant in well-being and personal wealth. Employability is directly related to the issue of professional qualification. According to various experts, the relationship between education level and income increases as the worker's educational level increases. Social inequality is the differentiation between people in the context of the same society. The main causes of inequality are lack of investment in social, cultural, health and education areas; mismanagement of resources, market logic, and ultimately; corruption. Epidemiological risk is defined as the probability of occurrence of a given health-related event, estimated from the event that occurred in the recent past. Thus, the risk of becoming a disabled person in a given population or group of people is the number of persons with disabilities occurring in the previous period by the number of persons existing in that period, since any person or all could potentially become disabled. - if persons with disabilities, who in turn have disadvantages in the labor market due to: lack of access to education, financing and training; urban and environmental barriers in the labor market, and finally; lack of perception by employers. Quality of life indicates the level of the basic and supplementary conditions of the human being. The race is understood as a social construct, used to distinguish people in terms of one or more socially significant physical markers responsible for creating a system of social inequality between different races, persons with disabilities and other marginalized categories. Income distribution represents the amount of money people receive from the amount of wealth they produce. Income concentration is the process by which income converges for the same company, region or group of people, unemployment, poverty, the growth of crime and violence. For this paper, it is considering data from the 2010 Population Census in the 5567 municipalities. As a statistical analysis of this work was considered: i) Multinomial logistic regression analysis; ii) Ordinal logistic regression stereotype; iii) annual profile diagram using compositional data, and finally step iv) the distribution across the 5565 municipalities.

**AUTHORS/INSTITUTIONS:** P.T. DE OLIVEIRA, EESC/STT, USP, São Paulo, SÃO PAULO, BRAZIL|

**CONTROL ID:** 3279926

**TITLE:** On a Modified g-Parameter Prior to Model Water Pollutants of Asejire River, Ibadan, Oyo State using Bayesian Model Averaging

**ABSTRACT BODY:**

**Abstract Body:** Bayesian Model Averaging (BMA) is a special technique that measures the uncertainties embedded in model selection processes, it depends on the appropriate model and parameter priors' choices. From the literature, the existing parameter priors based on fast increasing sample sizes compared to the number of regressors in a model give low Posterior Model Probability (PMP). In this research, an attempt was made to elicit a modified g -parameter priors to improve the performance of the PMP and predictive ability of the model with an application to the Water Pollutants' modelling of Asejire in Ibadan. The functional form of the modified g -parameter priors  $g_j =$  established the superiority of the consistency's conditions and asymptotic properties of the prior(s) using the Fernandez, Ley and Steel (FLS) models and with as sample sizes. The result of the analysis affirmed that the performance of PMP was reliable with the least standard deviations (0.1994 SD 0.0411) and (0.1086SD0:000) for model 1 and model 2 respectively; and it was convergent with the highest means (0.5378Mean0.9577) and (0.8342Mean1.000) for model 1 and model 2 respectively compared to the existing literatures. Additionally, the predictive performance affirmed the goodness of the elicited g-parameter priors when  $n = 50$  for Point prediction with (2.302, 2.357, 2.357); and when  $n = 100, 000$  for Overall prediction with (2.332, 2.334, 2.335) which were all closed to the LPS threshold 2.335 according to BMA specification. The implication of the analysis is that modified prior  $g_j =$  was a reliable parameter prior in the BMA based on this research and it was used to establish that dissolved solids (mg/l) and total solids (mg/l) were the most important pollutants in Asejire River of Ibadan, Oyo State with PIP of 6.14% and 6.1% respectively.

**AUTHORS/INSTITUTIONS:** S.A. Afolabi, Statistics, University of Ibadan, Ibadan, Oyo, NIGERIA|

**CONTROL ID:** 3280005

**TITLE:** Statistical challenges and opportunities using electronic health records for secondary data analysis

**ABSTRACT BODY:**

**Abstract Body:** The use of electronic health records (EHR) in research has become popular, as randomized clinical trials are time-consuming and costly. EHRs are an invaluable and rich source of data, but there are many challenges to overcome in order to utilize EHRs for clinical research. Clinically relevant information from the EHR permits the derivation of a rich collection of phenotypes. However, since EHR was primarily collected for clinical purposes, the true status of any given individual with respect to a certain trait of interest is not necessarily directly captured in the data set. In addition, although medication dosing information is critically important in many drug exposure-drug response related studies, an extracted dose information can be inaccurate, which needs to be addressed. The proposed abstract will include an overview of statistical methods and challenges of analyzing EHR, including estimation of latent phenotypes, development of algorithm for identification of correct medication dose, deriving and analyzing endpoints, and risk prediction. We demonstrate novel statistical approaches in real-world applications using EHR.

**AUTHORS/INSTITUTIONS:** J. Lee, Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, UNITED STATES|

**CONTROL ID:** 3281467

**TITLE:** Probabilistic modeling for an integrated temporary acquired immunity with norovirus epidemiological data

**ABSTRACT BODY:**

**Abstract Body:** Integration of acquired immunity into microbial risk assessment for illness incidence is of no doubt essential for the study of susceptibility to illness. In this study, a probabilistic model was set up as dose response for infection and a mathematical derivation was carried out by integrating immunity to obtain probability of illness models. Temporary acquire immunity from epidemiology studies which includes six different Norovirus transmission scenarios such as symptomatic individuals infectious, pre- and post-symptomatic infectiousness (low and high), innate genetic resistance, genogroup 2 type 4 and those with no immune boosting by asymptomatic infection were evaluated. Simulated results on illness inflation factor as a function of dose and exposure indicated that high frequency exposures had immense immunity build up even at high dose levels; hence minimized the probability of illness. Using Norovirus transmission dynamics data, results showed, and immunity included models had a reduction of  $2e6$  logs of magnitude difference in disease burden for both population and individual probable illness incidence. Additionally, the magnitude order of illness for each dose response remained largely the same for all transmission scenarios; symptomatic infectiousness and no immune boosting after asymptomatic infectiousness also remained the same throughout. With integration of epidemiological data on acquired immunity into the risk assessment, more realistic results were achieved signifying an overestimation of probable risk of illness when epidemiological immunity data are not included. This finding supported the call for rigorous integration of temporary acquired immunity in dose-response in all microbial risk assessments.

**AUTHORS/INSTITUTIONS:** E. Owusu-Ansah, Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, Ashanti, GHANA]

**CONTROL ID:** 3287390

**TITLE:** Modeling and analysis of Haldane genetic model in Brownian motion using the stochastic reaction-diffusion equation

**ABSTRACT BODY:**

**Abstract Body:** The mathematical modeling in genetics and exploring the links to physical disciplines are interesting because of their importance and beauties. In nature, the phenotypic traits have a complex dynamic inheritance, and are under the control of many genes and environment. The random fluctuations of the habitat change the frequency of gene in time and space,  $v(t,x)$ , so a suitable calculus is necessary for modeling the diffusion of  $v(t,x)$ . The aims of this paper are both to study the diffusion of  $v(t,x)$  with additive effects in a birth and death process of the Haldane genetic model using the Brownian motion model, and to evaluate the gene-environment interactions. It was found that if  $v(t,x)$  behaved like a super-Brownian motion and if the fatal mutations took place, then a tiny group of alleles were quickly disappeared. If  $v(t,x)$  was close to one, then it was not stable, and  $v(t,x)$  was tended to the stationary situation in the intermediate region according to the stabilizing selection conditions. There was not dominance effect, and if existed, it might be ambidirectional although the epistatic effect did not take place. The emerging of the dominance and epistatic effects were because of the directional selection. Another model was explained to study  $v(t,x)$  with white noise in the Falconer equations. The outlook can help model the similar problems in eco-evolutionary community genetics for studying the indirect genetic effects via the systems of stochastic partial differential equations, and white noise calculus.

**AUTHORS/INSTITUTIONS:** F. Fattahi, Plant breeding and biotechnology, Department of Environment of Kermanshah, Iran, Kermanshah, Kermanshah, IRAN (THE ISLAMIC REPUBLIC OF)

**CONTROL ID:** 3289329

**TITLE:** Early stopping in seamless phase I/II clinical trials

**ABSTRACT BODY:**

**Abstract Body:** In recent years, seamless phase I/II clinical trials have drawn much attention for dose finding taking into consideration both efficacy and toxicity as endpoints. Engaging an appropriate number of patients in a trial is always a challenging task. This paper proposes a dynamic stopping rule for seamless phase I/II clinical trials so that resources can be saved. That is, the stopping rule aims to employ a small number of patients if the optimum dose lies towards the beginning of a dose region. Similarly, the stopping rule aims to employ a large number of patients to identify the optimum dose accurately if it lies towards the upper end of a dose region. Particularly speaking, we allow a trial to stop early when width of the confidence intervals for the dose-response parameters become narrower or when the maximum sample size is exhausted, whichever comes first. Our simulation study of dose-response scenarios in various settings demonstrates that the proposed stopping rule can engage an appropriate number patients and therefore, we suggest its use in the clinical trials.

**AUTHORS/INSTITUTIONS:** M. Alam, N. Khan, University of Dhaka , Dhaka , BANGLADESH|

**CONTROL ID:** 3295066

**TITLE:** ANALYSIS OF THE EFFECT OF GREENHOUSE GAS EMISSION ON RAINFALL AND CORRESPONDING ENVIRONMENTAL OUTCOMES IN NIGERIA

**ABSTRACT BODY:**

**Abstract Body:** In this work, we investigated emission of Greenhouse gases (GHGs) in Nigeria from 1960 to 2014. Time series were extracted from the archives of the Worldbank on Global GHGs emission per country which contained records on Carbondioxide (CO<sub>2</sub>) from 1960-2014; Methane (CH<sub>4</sub>) and Nitrous-oxide (NO<sub>2</sub>) from 1970-2012. The Mann-Kendall trend test and Sen's slope were employed in establishing the presence of significant upward trend in the GHGs emission figures. Results showed that over the years CO<sub>2</sub> emissions rose by an average of 9.32% annually with a peak in 2005, CH<sub>4</sub> and NO<sub>2</sub> rose by an average of 3.30% and 2.58% respectively each reaching their peaks in 2007. Also, Rainfall records for 1971 to 2012 collected from the Nigerian Meteorological Agency (NIMET) was examined alongside GHGs emission quantities to ascertain the effect of the GHGs emissions on Climatic conditions such as Rainfall across climatic zones in Nigeria. Correlation results showed that significant relationship exist between the GHGs emission quantities and Annual rainfall in the Northern Guinea, Sudan and Sahel Savannah climatic zones. These results suggests the cause of the increase of 15.8%, 23.6% and 18.4% which were observed in average annual rainfall in the aforementioned climatic zones respectively since the 1990s -- which is one of the causes of incessant flooding experienced in the country since 2012 till date. Furthermore, using the double exponential smoothing method, predictions were made for the GHGs emissions up till 2050 which showed a steady increase in the emission figures which calls for immediate interventions in order to mitigate subsequent effects of climate change caused by these emissions. Regression analysis was employed to forecast annual rainfall amount based on the GHGs emission quantity forecast. Results showed that Annual rainfall in the Northern region is expected to increase in subsequent years which could lead to more devastating flood, loss of property, lives and a tremendous loss to the country.

**AUTHORS/INSTITUTIONS:** E.A. DOSUMU, O. OBISESAN, STATISTICS, UNIVERSITY OF IBADAN, Ibadan, Oyo, NIGERIA|

**CONTROL ID:** 3295225

**TITLE:** Modelling HIV/AIDS Disease Progression: A Parametric Semi-Markov Model with Interval Censoring

**ABSTRACT BODY:**

**Abstract Body:** Abstract

**Background:** HIV/AIDS epidemic continues to be the main killer disease in sub-Saharan Africa. The main objective of this study is finding factors affecting HIV/AIDS Disease Progression. This study was conducted to investigate the effect of factors on HIV/AIDS Disease Progression.

**Methods:** Patient follow-up data is obtained at Yirgalim General Hospital. A sample of 370 Patient data from a follow-up cohort is obtained at Yirgalem General Hospital. Multivariate generalized hazard regression model was employed to investigate the disease progression using both time independent and time dependent covariates.

**Results:** The study revealed that the risk of transition differs by patient's body mass index. Increase in the body mass index reduces the risk of transiting into the next worst states. The effects of sex, weight, age and body mass index of patients are significantly associated with AIDS disease progression. The risk of transition differs by patient's body mass index. Increase in the body mass index reduces the risk of transiting into the next worst states. The effect of sex, weight, age and body mass index of patients are significantly associated with AIDS disease progression. The results further revealed that the semi-Markov model with Weibull waiting time distribution has smaller log likelihood and AIC values compared to a semi-Markov model with exponential waiting time distribution.

**Conclusions and Recommendations:** In conclusion, transition probabilities are highly dependent on the choice of waiting times. We recommend that while choosing waiting time distributions for semi-Markov models one should consider appropriate distributions as waiting time distribution effect has a significant change on the estimated model parameters. In addition, this study recommends that concerned bodies should look at different contributing factors of AIDS diseases progression in addition to the ART services administered for slowing the current level of High diseased population in the country.

**AUTHORS/INSTITUTIONS:** T.F. Asena, Statistics, Arba Minch University, Arba Minch, ETHIOPIA|A.T. Goshu, Statistics, Kotebe Metropolitan University, Addis Ababa, ETHIOPIA|

**CONTROL ID:** 3295811

**TITLE:** IMPUTATION BASED ON A GAUSSIAN PROCESS FOR MISSING OBSERVATIONS OF FINE PARTICULATE MATTER

**ABSTRACT BODY:**

**Abstract Body:** Fine particulate matter ( $PM_{2.5}$ ) is one of the most aggressive air pollutants, responsible for many documented threats to human health and welfare. Indeed, Air Quality Monitoring Systems (AQMS) usually report hourly  $PM_{2.5}$  data, along with other pollution and climate variables. However, missing observations are not unusual, mostly due to common operation reasons like, for instance, power outages. On our research, we have found that, for  $PM_{2.5}$  measurements, there exists a Normal (univariate) distribution associated with data corresponding to hour  $i$  ( $i = 1, 2, \dots, 24$ ) of day  $j$  ( $j = 1, 2, \dots, 7$ ). Moreover, there also exists a joint Multivariate Normal distribution, with mean  $\mu_j$  y variance-covariance matrix  $\Sigma_j$ , for the 24 hours of day  $j$ . To see the whole setting, let's now assume we have complete data for a given year calendar. Then, we would have at least 52 observations (one per week) at each hour  $i$  of each day  $j$ . For example, we may have 52 observations at 10 am for day Thursday. However, on actual datasets, this number is usually below 52, due to missing data points for which we are proposing an imputation method. In all, we have 7 datasets, each one made of 24 subsets (one per hour) of  $PM_{2.5}$  observations. Each one of these 24 subsets would be of size at least 52, in case we had not missed any data point. In fact, we did have some complete days (those without missing points) in the dataset we used to illustrate the method, which led us to the usual framework on which functional data analysis (FDA) is conducted, with observations of a continuous phenomenon taken discretely. Thus, we may have as many as 52 curves per day each year calendar, which are assumed to be a sample from an infinite population. Since a Gaussian Process can be viewed as a class of models for which any finite-dimensional marginal distribution is Normal, then we used a Gaussian Process to generate the functional datum associated to day  $j$ . Then, we estimated the marginal normal distribution associated to hour  $i$  of day  $j$  at which we have a missing point. Finally, we used the estimated marginal Normal distribution to impute the missing data point at hour  $i$  of day  $j$ . We then did the same as needed for every missing data point and checked the properties of the imputation method. For illustration purposes we used  $PM_{2.5}$  data coming from the Cali (Colombia) AQMS.

**AUTHORS/INSTITUTIONS:** J. OLAYA, School of Statistics, Universidad del Valle, Cali, Valle del Cauca, COLOMBIA|

**CONTROL ID:** 3296680

**TITLE:** Bayesian Analysis of the Breast Feeding Pattern and Infant Mortality Rate in Nigeria

**ABSTRACT BODY:**

**Abstract Body:** Adequate nutrition is essential to children's growth and development. To this end, breastfeeding is universally endorsed by the World's health and scientific organizations as the best way of feeding infants, therefore early initiation of breastfeeding is important for both mother and child because inadequate nutrition can lead to infant mortality.

This study is aimed to give an update on the average number of last born children that were breastfed and to determine the rate of infant mortality across the geopolitical zones in Nigeria in 2008 and 2013.

Bayesian approach was used to analyse this problem. The lastborn children breastfed in Nigeria follows a Normal density and using a Normal prior yields a Normal Posterior. The infant mortality in Nigeria follows a Poisson distribution and using a conjugate Gamma prior yields a Gamma posterior. The average number of lastborn children breastfed (2008 and 2013 respectively) in North-Central (2473.676 and 1663.98), North-East (2678.324 and 2040.36), North-West (4903.023 and 4032.60), South-East (1697.717 and 1207.6), South-South (2295.461 and 1235.38) and South-West (2952.894 and 1685.03) were determined.

In 2008, the North-East zone had the highest infant mortality rate while the South-West zone had the lowest infant mortality rate while in 2013, the North-West zone had the highest infant mortality rate and the South-South zone had the lowest infant mortality rate.

It was recommended that there is need for more enlightenment on the part of Government and health agencies to nursing mothers on the relevance of breastfeeding to the growth of a child and the health of mothers.

**AUTHORS/INSTITUTIONS:** O.M. Oladoja, Statistics, University of Ibadan, Ibadan, Oyo, NIGERIA|

**CONTROL ID:** 3303178

**TITLE:** Principal Component Analysis of Coffee/Tea Drink Nutrition Science Research

**ABSTRACT BODY:**

**Abstract Body:** The purpose of this project is to determine which Starbucks drinks among all coffee and tea options are best for cardiovascular disease (CVD) prevention and overall good health. A Science-based Health Index was constructed considering different coffee/tea constituents, including: saturated fat, cholesterol, sodium, carbohydrates, dietary fiber, sugars, protein, and caffeine. Antioxidant activity of flavonoids from Caffeine contained within Coffee/Tea can reduce free radical formation and scavenge free radicals. Principal Components Analysis (PCA) was used to explore all factors in the analysis and to inform on the utility of the health index in relation to its link to CVD prevention and good health. Principal Component 1 is more relevant to most unhealthy components such as sugars, carbohydrates, saturated fat, and total fat. Principal Component 2 is more related to healthy Caffeine nutrition. Additionally, Dietary Fiber and Caffeine are most opposite against the other unhealthy components along the direction of the both 1<sup>st</sup> and 2<sup>nd</sup> Principle Components. PCA Eigen Analysis is very powerful computational and visual diagnostic tool for discrimination and classification of coffee product types based on patterns in nutritional constituents. To avoid variance factor in PCA analysis, original data has been Z-transformed and JMP loadings plots are standardized. The new PCA-based Health Index was derived based on the eigenvalues and eigenvectors of the first two Principle Components. The new PCA-based Health Index was also compared and correlated to the Science-based Health Index (about 70%-80% R-Square Curve Fitting). Due to the orthogonality of Principle Eigen analysis, the remaining eight principle components are neutral on the Health Index (~0% R-Square). The Principle component analysis has also demonstrated Pareto Concept (the first 20% Principle Components has addressed the 79% Variance). The other Foods like Candy, Chocolate, Cereal are also studied and their Nutrition Loadings plots of the first two principle components are quite different each other. The loadings plot pattern can provide insight indication of product nutrition and health information.

**AUTHORS/INSTITUTIONS:** C. Chen, Continuous Improvement, Applied Materials, San Jose, California, UNITED STATES|P. Giuliano, Quality Engineering, Abbott, Menlo Park, California, UNITED STATES|M. chen, OHS, Stanford, Palo Alto, California, UNITED STATES|

**CONTROL ID:** 3303208

**TITLE:** Study Body Reaction of Altitude Sickness and Fatigue Research

**ABSTRACT BODY:**

**Abstract Body:** This paper will address the risk of suffering the Altitude Sickness when people are hiking on the high Mountains without caution. It's difficult for most people to accommodate the high-Altitude environment suddenly. It's very risky if the people are not aware of their altitude sickness symptom such as Fatigue, Headache, Dizziness, Insomnia, Shortness of breath during exertion, Nausea, Decreased appetite. The consequence of certain altitude sickness could be in very dangerous situation on the inconvenient high mountains. There are three kinds of altitude sickness: (1) Acute Mountain Sickness (AMS) is the mildest form and it's very common, (2) High Altitude Pulmonary Edema (HAPE) is a buildup of fluid in the lungs that can be very dangerous and even life threatening, and (3) High Altitude Cerebral Edema (HACE) is the most severe form of altitude sickness and happens when there's fluid in the brain. It's life threatening, and you need to seek medical attention right away. To detect the Altitude Sickness in real time, a Pulse Oximeter was used to monitor the Oxygen% and Heart Beat at different altitude levels from near sea level in San Jose to Denver (5,000 Feet), Estes Park (8,000 Feet), Rocky Mountains Alpine Center (> 12,000 Feet). Light to medium Altitude Sickness was observed when reaching the Altitude height over 8,000-12,000 feet. To simulate the hiking and climbing risk, at 12,000 feet, a 2.5mins Jumping Rope exercise was conducted to analyze the fatigue behavior and impact to the Altitude Sickness. Statistical analysis was conducted to verify several hypotheses to predict how high of the Altitude Sickness Risk at different altitude levels as well as the additional Exercise Fatigue Behavior. This method may provide the people how to assess their body strength and readiness before they may take a long hiking on the high mountains. The experimental results are showing that both the heart beat and Oxygen level are sensitive to the Altitude Sickness degree but only the heart beat could detect the fatigue status during the Jumping Rope exercise.

**AUTHORS/INSTITUTIONS:** C. Chen, Continuous Improvement, Applied Materials, San Jose, California, UNITED STATES|M. chen, OHS, Stanford, Palo Alto, California, UNITED STATES|

**CONTROL ID:** 3313702

**TITLE:** A Three-Parameter Gompertz-Lindley Distribution: Its Properties and Applications

**ABSTRACT BODY:**

**Abstract Body:** This research proposed a three-parameter probability distribution called Gompertz-Lindley distribution that can be used for modelling survival analysis data using Gompertz generalized (Gompertz-G) family of distributions. The mathematical properties of the distribution such as moment, moment generating function, survival function and hazard function were derived. The parameters of the distribution were estimated using the method of maximum likelihood and the distribution was applied to model the strength of glass fibres. Gompertz-Lindley distribution performed best (AIC = 62.8537) when compared with other generalizations of the Lindley distribution to model real life data.

**AUTHORS/INSTITUTIONS:** P.O. Koleoso, Statistics, University of Ibadan, Ibadan, Oyo State, Ibadan, Oyo State, NIGERIA|

**CONTROL ID:** 3331068

**TITLE:** Zero-inated Defective Regression Models for Cure Rate Modeling

**ABSTRACT BODY:**

**Abstract Body:** In this work, we introduce a zero-inflated defective regression model. Our approach enables us to accommodate three types of patients, that is, patients with zero survival time (failure occurred at the moment the study began) and those who are susceptible or not susceptible to the event of interest. An advantage of our proposal is that it accommodates zero inflated lifetimes, which is not possible in the standard defective models. Defective distributions are obtained from standard distributions by changing the domain of the parameters of the latter in such a way that their survival functions are limited to  $(0; 1)$ . We consider the Gompertz defective distributions, which allow modeling of data containing a cure fraction. Parameter estimation is performed by maximum likelihood estimation, and Monte Carlo simulation studies are conducted to evaluate the performance of the proposed models. We illustrate the practical relevance of the proposed models on real data sets on insulin use in pregnant women diagnosed with gestational diabetes performed at São Paulo University Medical School.

**AUTHORS/INSTITUTIONS:** V.L. Tomazella, Estatística, Universidade Federal de São Carlos, São Carlos, São Paulo, BRAZIL|V.F. Calsavara, Department of Epidemiology and Statistics, 1A.C.Camargo Cancer Center, São Paulo, São Paulo, BRAZIL|A. Rodrigues, Estatística, Universidade Federal do Espírito Santo, Vitória, Espírito Santo, BRAZIL|R. Rocha, Department of Statistics,, Universidade Feral da Bahia, Salvador, Bahia, BRAZIL|

**CONTROL ID:** 3331224

**TITLE:** Sequential imputation under stereotype models with application to cancer registry data

**ABSTRACT BODY:**

**Abstract Body:** Incomplete categorical data analysis presents unique challenges. We present an adaptation of variable-by-variable imputation using a new class of models for ordinal data. Specifically, we consider stereotype models for the ordinal variables and combine them with computational efficient algorithms for continuous variables. Resulting set of models and computational algorithms easily overcome the burden of dimensionality specific to categorical data. Further, by definition, the variable-by-variable imputation approach solves the typical complexities stemming from skip patterns, bounds and restriction as commonly seen in surveys. We present a comprehensive simulation study on the performance of our proposed methods which are also illustrated using cancer registry data.

**AUTHORS/INSTITUTIONS:** R. Yucel, Epidemiology and Biostatistics, SUNY-Albany, Rensselaer, New York, UNITED STATES|D. Fernández, Institut de Recerca, Parc Sanitari Sant Joan de Déu, Barcelona, Barcelona, SPAIN|

**CONTROL ID:** 3338132

**TITLE:** Prediction of Failure Times of Censored Items for a Simple Step-Stress Model with Hybrid Censoring from the Exponential Distribution

**ABSTRACT BODY:**

**Abstract Body:** In this article, the problem of predicting times to failure of units from the Exponential Distribution which are censored under a simple step-stress model is considered. We discuss two kinds of predictors - the maximum likelihood predictors (MLP) and the conditional median predictors (CMP) in the context of Type I and Type II hybrid censoring scheme (HCS). In order to illustrate the prediction methods we use some numerical examples.

Furthermore, mean squared prediction error (MSPE) and prediction intervals are generated for these examples using simulation studies. MLP and the CMP are then compared with respect to the MSPE and prediction interval for each type of censoring. Finally, we used a real data to apply the prediction methods developed in the article.

**AUTHORS/INSTITUTIONS:** I. Basak, Math/Stat, Penn State Altoona, Altoona, Pennsylvania, UNITED STATES|

**CONTROL ID:** 3338277

**TITLE:** Weight Calibration to Improve the Efficiency of Pure Risk Estimates from Case-control Samples Nested in a Cohort

**ABSTRACT BODY:**

**Abstract Body:** Cohort studies provide information on relative hazards and pure absolute risks of disease. For rare outcomes, large cohorts are needed to have sufficient numbers of events, making it costly to obtain covariate information on all cohort members. We focus on nested case-control designs that are used to estimate relative hazards in the Cox regression model. Langholz and Borgan (1997) showed that pure risk (survival probability) can also be estimated from nested case-control data. However, these approaches do not take advantage of some covariates that may be available on all cohort members. Researchers have used weight calibration to increase the efficiency of relative hazard estimates from case-cohort studies and nested case-control studies. Our objective is to extend weight calibration approaches to nested case-control designs to improve the precision of estimates of relative hazards and pure risks. We show that calibrating sample weights additionally against time-on-study information (follow-up times during the risk projection period) improves estimates of pure risk. Efficiency improvements for relative hazards for variables that are available on the entire cohort also contribute to improved efficiency for pure risks. We develop explicit variance formulas for the weight-calibrated estimates. Simulations show how much precision is improved by calibration and confirm the validity of inference based on asymptotic normality. Examples are provided using data from the American Association of Retired Persons (AARP) Diet and Health Cohort Study.

**AUTHORS/INSTITUTIONS:** Y. Shin, R. Pfeiffer, B. Graubard, M. Gail, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, UNITED STATES|

**CONTROL ID:** 3338817

**TITLE:** An evaluation of safe water programs serving people living with HIV/AIDS, Ethiopia, 2008

**ABSTRACT BODY:**

**Abstract Body:** Background: Diarrhea is a leading cause of illness and death among people living with HIV/AIDS (PLWHA) in sub-Saharan Africa, and household water chlorination has been shown to reduce diarrhea incidence among PLWHA. In the past, some HIV programs in Ethiopia provided a household-based drinking water quality intervention (brand name WuhaAgar), a socially marketed chlorination product, to prevent diarrhea.

Methods: To evaluate the water program, we compared WuhaAgar use and water treatment practices between 795 clients from 20 antiretroviral treatment (ART) clinics (ART clinic clients) and 795 community members matched by age, sex, and neighborhood. Univariable conditional logistic regression was used to assess associations between ART clinic clients and selected covariables and calculate matched odds ratios (mOR).

Results: Overall, 19% of study participants reported water treatment with WuhaAgar. There was a positive association between ART clinic clients and use of improved water storage (mOR: 2.7 95% CI: 1.7-4.3), reported drinking water treatment (mOR: 3.8 95% CI: 2.9-5.0), reported current water treatment with WuhaAgar (mOR: 5.5, 95% CI 3.9-7.7), and bottles of WuhaAgar observed in the home (mOR: 8.8, 95% CI 5.4-14.3). There was also a positive association between ART clinic clients and reported diarrhea among respondents (mOR: 4.8, 95% CI 2.9-7.9) and household members (mOR:2.8, 95% CI: 1.9-4.2) in the two weeks preceding the survey. These associations were stronger in clients who attended ART clinics that stocked WuhaAgar than in clients of ART clinics in healthcare facilities that did not stock WuhaAgar. The strength of association between clients and diarrhea was unchanged when current water treatment was adjusted in the model.

Conclusion: Study results suggest that promoting and distributing water chlorination products in ART clinics was effective in increasing access to and reported use of water treatment products among PLWHA. The positive association between ART clinic attendees and diarrhea likely resulted from the immunocompromised status of ART clinic clients.

**AUTHORS/INSTITUTIONS:** S. kim, C. Reilly, Z. Salah, A. Bhattarai, R. Quick, Centers for Disease Control and Prevention, Atlanta, Georgia, UNITED STATES|

**CONTROL ID:** 3340305

**TITLE:** Mixture of Linear Mixed Models for Clustering Gene Expression Profiles from Repeated Microarray Experiments

**ABSTRACT BODY:**

**Abstract Body:** Data variability can be important in microarray data analysis. Thus, when clustering gene expression profiles, it could be judicious to make use of repeated data. In situations where a large data set is partitioned into many relatively small groups, and where the members within a group have some common unmeasured characteristics, the number of parameters requiring estimation tends to increase with sample size if a fixed effects model is applied. This fact causes the assumptions underlying asymptotic results to be violated.

A method is proposed that aims at identifying clusters of individuals that show similar patterns when observed repeatedly. We considered linear mixed models which are widely used for the modeling of longitudinal data and to also take into account data variability and mixture of these models. In contrast to the classical assumption of a normal distribution for the random effects, a finite mixture of normal distributions is assumed. Typically, the number of mixture components is unknown and has to be chosen, ideally by data driven tools. This leads to a large range of possible models depending on the assumptions made on both the covariance structure of the observations and the mixture model.

Here, we considered two possible solutions to this problem, a random intercepts model and a fixed effects model, where asymptotic are replaced by a simple form of bootstrapping. A profiling approach is introduced in the fixed effects case, which makes it computationally efficient even with a huge number of groups.

A simulation studies is considered and properties of the proposed statistic, are investigated and also applied to real life data of Breast Cancer from Nigeria.

**AUTHORS/INSTITUTIONS:** V.E. Laoye, Statistics, University of Ibadan, Nigeria, Ibadan, NIGERIA|

**CONTROL ID:** 3340737

**TITLE:**

Determinants of Isoniazid Preventive Therapy Completion among People Living with HIV Attended Care and Treatment from 2013 to 2017 in Dar es Salaam Region, Tanzania. A cross-sectional analytical study

**ABSTRACT BODY:**

**Abstract Body:** Abstract

**Background:** Tuberculosis (TB) disease is a common opportunistic infection among people living with HIV (PLHIV). WHO recommends at least six months of isoniazid Preventive Therapy (IPT) to reduce the risk of active TB. It is important to monitor completion of IPT, as a suboptimal dose may not protect PLHIV from TB. This study determined IPT completion and its determinants among PLHIV in Dar es Salaam region, Tanzania.

**Methods:** A Cross-sectional analytical study was conducted using secondary analysis of routine data from 58 care and treatment clinics in Dar es Salaam region. The study recruited clients who screened negative for TB symptoms and initiated IPT between January 2013 and June 2017. A modified Poisson regression model with robust standard errors was used to estimate prevalence ratios (PR) and 95% confidence interval (CI) for factors associated with IPT completion.

**Results:** A total of 29382 clients were initiated on IPT, with 21808 (74%) female. Overall 17092 (58%) completed IPT, increasing from 42% (773/1857) in year 2013 to 76% (2929/3856) in 2017. There was lower IPT completion among those were not on ART (PR: 0.56: 95%CI: 0.49-0.64); those with advanced WHO clinical stage III (PR: 0.94: 95%CI: 0.92-0.97) and IV (PR: 0.93: 95%CI: 0.89-0.98). Compared to PLHIV with CD4 counts less than 100 cells/, those with CD4 counts between 100 and 349 cells/ had higher IPT completion (PR: 1.03: 95%CI: 1.01-1.05) and those with high CD4 counts  $\geq 350$  cells/ had lower IPT completion (PR: 0.95: 95%CI: 0.93-0.99).

**Conclusion:** IPT completion prevalence is low at care and treatment clinics although it increased over time. The lower IPT completion seen in PLHIV with advanced-stage HIV disease and those, not on ART indicates the need for better IPT interventions with greater support PLHIV in those groups.

**Keywords:** IPT completion, HIV, tuberculosis, predictors, Tanzania

**AUTHORS/INSTITUTIONS:** M. Robert, Undergraduate science, Mwenge Catholic University, Moshi, Kilimanjaro, TANZANIA, UNITED REPUBLIC OF|M. Robert, Biostatistics, Kilimanjaro Christian Medical University College, Moshi, Kilimanjaro, TANZANIA, UNITED REPUBLIC OF|

**CONTROL ID:** 3340765

**TITLE:** Improved models for analyzing zero-inflated data with application on pineapple (*Ananas Comosus* L. Merr.) data containing excess zeros

**ABSTRACT BODY:**

**Abstract Body:** In agriculture, count data often contain a large number of zero observations. Sample size requirements are also common. Pineapple is a tropical plant with edible multiple fruit consisting of coalesced berries and the most economically significant plant in the Bromeliaceae family. It is subject to a variety of diseases and the most serious is wilt disease vectored by mealybugs typically found on the surface of pineapples. Symptoms include yellowing and wilting of the leaves, resulting in the death of infected plants. Field data from experiment carried out along climatic zones in Benin revealed excess number of zeros for wilted plants across cultivars of pineapples. More than 50 % of the observations were zero. Failing to take excess zeros into account in modeling process can lead researchers to erroneously conclude. In this study we propose a new modeling method to overcome the problems caused by zero-inflated data sets. We combined zero inflated models, which are widely used in agriculture, and Random Forests (RF) model, a machine-learning technique. The study assessed empirically, the performance of zero inflated models and RF model under varying proportion of zero observations and sample size using R software. Increased values of zero observations ( $p=0.20, 0.40, 0.60, 0.80$  and  $0.90$ ) have been introduced in count data and various samples ( $n=25, 50, 100, 500, 1000$ ) were extracted. Poisson, negative binomial, zero inflated Poisson and zero inflated negative binomial models were fitted on the sampled data and the results were compared to the outcomes from RF model performed on the same data. The models were compared in terms of mean bias and mean squared error. The models were then applied on real life data collected on prevalence of wilt disease on the pineapple cultivars in Benin. Results showed that sample size and proportion of zero observations impact the performance of the models studied. Random Forests model was better than the other models for large samples ( $n=500$  and  $1000$ ) and highest proportion of zero ( $0.60, 0.80$  and  $0.90$ ). However, for small samples ( $n=25, 50$ ) and small  $p$  ( $p=0.2, 0.4$ ) there was no model performing better in all conditions. The application on pineapple data showed that RF model performed better than other models. RF model is a promising method for analyzing zero-inflated data sets in agriculture.

**AUTHORS/INSTITUTIONS:** B.E. Lokonon, Laboratoire de Biomathématiques et d'Estimations Forestières, Université d'Abomey-Calavi (Benin), Abomey-Calavi, BENIN|

**CONTROL ID:** 3341208

**TITLE:** Using penalized methods for propensity score model with rare exposure in observation studies with binary outcome

**ABSTRACT BODY:**

**Abstract Body:** Propensity score methods are increasingly being used for estimating causal effects of exposure on outcome in the presence of confounders in observational studies. In such analysis, the propensity scores (PS) are usually estimated as the predicted probability of being exposed using the logit or probit model of exposure with the baseline covariates. The PS are then used in matching, weighting, and covariate-adjustment for estimating marginal effect of the exposure to make causal inference. All these approaches perform well when the prevalence of exposure is reasonably high. However, there is doubt in the performance of different PS methods when exposure is rare in practice. Because, like rare outcome model, maximum-likelihood (ML) based logit or probit model yield biased estimate of the predicted probability of being exposed (i.e., PS) for rare exposure, which in turn may affect the estimate of the marginal effect in the outcome model and make misleading casual inference. This research proposed penalized methods for PS models for achieving accurate estimate of the predicted probability (i.e., PS) when exposure is rare. We explored the use of both Jeffreys-and log F(1,1)-prior based penalized logistic regression for estimating the PS and their use in PS matching, weighting, and covariate-adjustment to examine if the penalized PS improve the performance of estimating accurate casual effects over the standard PS. Extensive simulation study was performed by creating several scenarios reflecting rare exposure. Simulation results revealed that all penalized methods showed improvement to some extent, in terms of bias and MSE associated with the marginal estimate (odds ratio) in the outcome model, over the standard approach, which is true for all type of PS analyses. Of them, log F(1,1) prior based penalized method showed comparatively greater improvement. Further, the methods were illustrated using rare exposure data for estimating casual effects of malnutrition on acute respiratory infection in children of age under five years.

**AUTHORS/INSTITUTIONS:** M.S. Rahman, T. Rahman, Institute of Statistical Research and Training, University of Dhaka, Dhaka, BANGLADESH|

**CONTROL ID:** 3341320

**TITLE:** Prioritizing Disease Candidate Genes Using Knockout Mouse Phenotype Data

**ABSTRACT BODY:**

**Abstract Body:** To characterize the relationship between protein-coding genes and phenotypes, the International Mouse Phenotyping Consortium (IMPC) is creating an extensive catalogue of mammalian gene function by i) producing knockout mouse lines for the approximately 20,000 protein-coding genes, ii) conducting systemic phenotyping on every knockout line and iii) studying the association between gene-knockout and phenotype. This unique and comprehensive open data set of genome and phenome-wide association opens an unprecedented opportunity to understand the etiology and underlying mechanism of diverse human diseases/traits. As a proof of concept of its utility in human disease studies, here we show that the IMPC gene-to-phenotype association data can be utilized to prioritize candidate genes in genome-wide association studies (GWAS) of psychiatric disorder (e.g., PGC Schizophrenia GWAS). In particular, we identify and prioritize genetic loci/genes associated with sensorimotor gating deficits in schizophrenia by using synthetic gene scores obtained from a meta-analysis combining association across multiple sensory gating phenotypes available in IMPC data (e.g., prepulse inhibition).

**AUTHORS/INSTITUTIONS:** D. Lee, Statistics, Miami University, Oxford, Ohio, UNITED STATES|

**CONTROL ID:** 3342326

**TITLE:** Roles of Frailty in Modelling Competing-Risks Data and Extensions: H-likelihood Framework

**ABSTRACT BODY:**

**Abstract Body:** Frailty, the effect of unobserved random covariates on the risk of a patient, is very useful in modelling heterogeneity and/or dependence among time-to-event data. Until now, semi-parametric frailty models with unspecified baseline hazard have been widely used for the analysis of various survival data; particularly, ignoring this frailty can lead to biased estimate of treatment effect. However, for the model inference the frailty term has been often regarded as a nuisance, leading that it does not provide directly inference on frailty due to elimination of frailty term by integration. In this talk, we introduce various roles of frailty in analyzing competing risks (CR) data allowing for different types of events which can be correlated, via h-likelihood (Lee and Nelder, 1996; Ha et al., 2017). Unlike the classical likelihood for fixed parameters only, the h-likelihood is constructed for both fixed parameters and unobserved frailties at the same time, and it provides efficient statistical inference for various univariate and multivariate survival models, together with frailtyHL R package (Ha et al., 2017). For the purpose, we consider to add frailty term into two different but popular CR models, i.e. cause-specific and sub-distribution hazard models. In particular, with inference of individual frailties in both CR models we show how to investigate the heterogeneity of treatment effect across centers (i.e. random treatment-by-center interaction) in multi-center clinical trial under competing risks setting. Furthermore, we present the usefulness of frailty in modelling semi-competing risks data where only a terminal event (e.g. death) censors a non-terminal event (e.g. disease recurrence); a patient may experience both events that may be correlated. We also discuss about extended CR models allowing for frailties in another survival models including copula models, accelerated failure time models, joint models and mean residual life models.

**AUTHORS/INSTITUTIONS:** I. Ha, Pukyong National University, Busan, KOREA (THE REPUBLIC OF)|Y. Lee, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3342548

**TITLE:** A new method for multivariate functional data classification, with application to detecting difficulty in computer-assisted surveys through mouse movement trajectories

**ABSTRACT BODY:**

**Abstract Body:** One of the main goals of surveys is to collect robust and reliable data from respondents and conversely to reduce sources of error. One source of error stems from participants' difficulty in understanding and responding to survey questions in the way the researchers intended, and thus detecting and mitigating issues promises to improve both the respondent experience and data quality (Horwitz et al. 2017). In the presence of a human interviewer, difficulty can be assessed by identifying and quantifying paralinguistic cues, etc. In web surveys, these difficulties can be evaluated using auxiliary survey data that describe the answering process, also called paradata (e.g., mouse clicks and movements).

This work aims to identify instances when online survey respondents experience difficulty and confusion while answering by comparing and classifying their mouse movement trajectories. In an online survey, we experimentally manipulated the difficulty of questions related to participants' current employment status and general personal situation (e.g., using concise vs. complex language). To predict whether respondents faced the easy vs. complex version of the question only from their mouse movement trajectories, we developed a method to classify multivariate functional data, combining and extending the semi-metric based techniques introduced by Fuchs et al. (2015) and Ferraty and Vieu (2003). We go beyond existing methods both by the extension to the multivariate functional setting as well as by introducing a personalization method to control for the baseline mouse behavior of the survey participants. We demonstrate that accounting for individual differences is critical to successful prediction of difficulty. An R package implements the presented classification methods for multivariate functional data and trajectories in  $n$  dimensions.

Horwitz R, Kreuter F, Conrad F. Using Mouse Movements to Predict Web Survey Response Difficulty. *Social Science Computer Review* (2017), 35(3): 388-405

Fuchs K, Gertheiss J, Tutz G. Nearest neighbor ensembles for functional data with interpretable feature selection. *Chemometrics and Intelligent Laboratory Systems* (2015), 146: 186-197.

Ferraty F, Vieu P. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis* (2003), 44: 161-173.

**AUTHORS/INSTITUTIONS:** A. Fernández-Fontelo, S. Greven, Humboldt University of Berlin, Berlin, GERMANY|A. Fernández-Fontelo, Departament de Matemàtiques, Universitat Autònoma de Barcelona, Barcelona, SPAIN|F. Henninger, P. J. Kieslich, Mannheim Centre for European Social Research, University of Mannheim, Mannheim, GERMANY|F. Kreuter, University of Maryland, University of Maryland, College Park, Maryland, UNITED STATES|

**CONTROL ID:** 3343107

**TITLE:** Silicon alleviates arsenic toxicity: evidences based on Krebs cycle, GABA and polyamine synthesis in rice

**ABSTRACT BODY:**

**Abstract Body:** Arsenic toxicity is a global concern due to ever-increasing groundwater contamination, crops irrigation in many regions of the world. Arsenic exists in two important oxidation states – arsenate and arsenite. The purpose of the investigation was to determine ameliorative effects of silicon on Krebs cycle,  $\gamma$ -aminobutyric acid and polyamine synthesis in rice (*Oryza sativa* L. cv. MTU-1010) seedlings under arsenic stress. Arsenate is reduced to arsenite by arsenate reductase that leads to decrease in growth and water contents in arsenate treated rice seedlings. Silicate application in arsenate treated seedlings altered these effects significantly according to Tukey's HSD multiple comparison analysis. In the test seedlings under arsenate application, the activities of respiratory enzymes viz., pyruvate dehydrogenase, isocitrate dehydrogenase,  $\alpha$ -ketoglutarate dehydrogenase, succinate dehydrogenase, fumarase, malate dehydrogenase and citrate synthase were decreased while the levels of organic acids viz., pyruvate, citrate, succinate and malate were increased. But joint application of silicate along with arsenate increased the activities of all mentioned respiratory enzymes resulted in more elevation of organic acid contents in the seedlings of test cultivar.  $\gamma$ -aminobutyric acid accumulation along with the activities of its regulatory enzymes viz., glutamate dehydrogenase and glutamate carboxylase were enhanced under arsenate stress while under co-application with arsenate and silicate, the said parameters were down-regulated indicating a protective role of silicon against arsenic stress in rice seedlings. Polyamines have the ability to generate tolerance against various environmental stresses in plants. In the test seedlings the polyamines viz., putrescine, spermidine and spermine were synthesized more during joint application of arsenate with silicate. Therefore, silicate supplementation substantially moderated the toxic effects of arsenate in the test cultivar. Thus, application of silicon in arsenic contaminated soil may be a useful approach to grow rice by alleviating stress effects as well as resulting in the revival of potential health risks.

**AUTHORS/INSTITUTIONS:** S. Das, A. Biswas, Botany, University of Calcutta, Kolkata, INDIA|

**CONTROL ID:** 3343445

**TITLE:** Modelling Language Extinction Using Susceptible-Infectious-Removed (SIR) Model

**ABSTRACT BODY:**

**Abstract Body:** A stochastic epidemic model, applied to model indigenous language extinction is presented in this study. The Susceptible-Infectious-Removed (SIR) categorization of an endemic disease is reformulated to capture the dynamics of indigenous language decline based on the assumption of non-homogeneous mixing. The expected time to extinction of the indigenous language is derived using a modified SIR model with the population segmented into several sub-communities of small sizes representing the family units. Data obtained from an indigenous language survey conducted in some parts of Nigeria and the Nigeria Demographic Health Survey (NDHS) were used to estimate key parameters of the model for some of Nigeria's indigenous languages. The parameters of interest included the basic reproduction number, the threshold of endemicity and the expected time to extinction, starting from the endemic level. On the basis of the expected time to extinction, several of the surveyed languages were seen to be in a precarious condition while a few others were seemingly virile based on a high language transfer quotient within families.

**AUTHORS/INSTITUTIONS:** N. Ikoba, E.T. Jolayemi, Statistics, University of Ilorin, Ilorin, Kwara, NIGERIA|

**CONTROL ID:** 3343773

**TITLE:** Quantification of Uncertainty for Data-driven Survival Projection in Cancer

**ABSTRACT BODY:**

**Abstract Body:** Background

Cancer survival trend projection is crucial to monitor patient's prognosis at the population-level and evaluate success against cancer control. Data driven survival projection models such as joinpoint survival model, flexible parametric model and period analysis extrapolate current trends to project unknown survival in the future. All survival models are based on the specified assumptions and observed data, and the reliability of the projection depends on the uncertainties in the models.

Objectives

The aim of this study is to compare survival trend projection models and assess level of the uncertainties in data-driven survival projection based on the population-based cancer registry data.

Methodology

Survival from the Korean lung cancer patients (1993-2016) were analyzed and projected into the future. Three different survival trend projection models were compared: joinpoint regression models (JP) to project survival, identifying trends with significant change points in survival; flexible parametric survival models (FP) to adequately estimate complex patients' underlying risk; period analysis (PA) to estimate up-to-date survival focusing on the recent data. We characterized the possible locations of uncertainties including the model structures, inputs, model outcome, and the survivals were projected based on the different length of period to quantify the uncertainty level. The mean squared errors (MSE) of the survival estimates were computed to evaluate accumulated uncertainties.

Result

Individual records were used for projection based on FP and PA, while JP was fitted using collapsed data. The estimated the five-year relative survival for patients with lung cancer and projected the survivals by 2030 were 42.3%(JP), 36.5%(FP), and 32.2%(PA) in male. PA provided the higher estimates compared to JP for localized and regional stage, reflecting improved survival in recent period. The projected survival from JP showed 9.3% differences as input data were accumulated since 2010. MSE were higher for PA compared to JP and FP since the improved survival in recent period back extrapolated the survival in the past.

Conclusion

This study demonstrated uncertainties of future survival estimates based on the different survival projection models and showed the presence of nonnegligible uncertainties, indicating a need to increase the reliability of model by updating uncertain parameters.

**AUTHORS/INSTITUTIONS:** S. Lee, H. Cho, Cancer Control and Population Health, National Cancer Center, Goyang-si, Gyeonggi-do, KOREA (THE REPUBLIC OF)

**CONTROL ID:** 3344069

**TITLE:** Recent statistical challenges in analysis of copy number alterations from next-generation sequencing

**ABSTRACT BODY:**

**Abstract Body:** Next-generation sequencing (NGS) has transformed and improved our investigation of genomic profiles such as copy number alteration (CNA). Copy number alterations are genomic regions that exhibit more ("gains") or less ("losses") copies than the normal two copies. However, there are many steps need to be taken to make sense the information contained in the sequence data. It starts from the moment we obtain mapped sequence data to the use of the data for clinical subtyping or identification of important genomic regions. In each and every steps involved, statistical modelling and analysis are needed to address those challenges. In this talk, I will discuss and present the challenges and the statistical methods to address those challenges. This includes segmentation and normalisation to estimate CNA, modelling strategies based on CNA profile, and developing new statistical test per genomic region. This is joint work with Henry Wood, Stefano Berri, Pamela Rabbitts, Mohammed Alshahrani, Khaled Alqahtani, Alison Telford, and Charles Taylor.

**AUTHORS/INSTITUTIONS:** A. Gusnanto, Statistics, University of Leeds, Leeds, UNITED KINGDOM|

**CONTROL ID:** 3344322

**TITLE:** ESTIMATION OF AGE DEPENDENT EXPECTED ACCOMPLISHMENT PROBABILITIES IN A POPULATION

**ABSTRACT BODY:**

**Abstract Body:** This paper proposes and present expected accomplishment probability model during each of the three age epochs in a population. Methods for the estimation of these probabilities are also proposed and presented. Test statistic are proposed and presented for use in testing and desired hypotheses about expected accomplishment probabilities. The proposed methods are illustrated with some sample data on retirees from a certain population in which human life or age span is partitioned into three, namely, less than 25 years, 25 to 50 years, and over 50 years. The results obtained using these sample data showed that less than one-fourth of retirees from the sampled population are able to accomplish all that is normally expected of subjects from the population after retirement from relatively active work. However, only about one-fifth of retirees seem unable to accomplish all that is normally expected following retirement.

**AUTHORS/INSTITUTIONS:** C.A. UZUKE, STATISTICS, NNAMDI AZIKIWENUNIVERSITY AWKA, Awka, Anambra, NIGERIA|

**CONTROL ID:** 3344479

**TITLE:** Using A Behavior Change Techniques Taxonomy to Identify Active Ingredients within Interventions for Improving the appropriate use of polypharmacy for older people: A Meta-regression Combining Summaries of Binary Outcomes with Those of Count Outcomes

**ABSTRACT BODY:**

**Abstract Body:** A meta-analysis should ideally consist of all trials ever conducted to investigate a specific research question. An application of meta-regression analysis will be presented that combines following trials: trials that reported continuous outcome measures, trials that reported binary outcomes created by a dichotomy of the count outcome and other trials that reported both of count and binary outcomes.

This was motivated by a series of controlled clinical trials investigating the effect of complex interventions to prescription behavior of physicians to prevent aged patients from potentially inappropriate medications. The difference in the magnitude of the effect was assumed to be due to the difference in behavior techniques used and techniques contributing to the magnitude of the effect were searched.

Outcome measures were obtained in the form of summary statistics from published papers. Behavior change techniques that were used in the complex interventions were identified according to behavior change technique taxonomy version 1 (BCTTv1) by analyzing description of papers. Primarily, the log-odds ratio was used as a common measure of intervention difference across all trials.

Compared to the meta-regression analysis that used only 11 trials which report binary outcomes, the meta-regression analysis that used 15 trials showed different significant behavior change techniques related to improvement in proportion of potentially inappropriate medications.

By choosing a measure of intervention difference such as the log-odds ratio, which can be estimated from both binary and count data, the number of trials included in the meta-analysis can be increased and better representation ensured. The results and the impact of different model assumptions to the number of significant behavior change techniques will be presented.

**AUTHORS/INSTITUTIONS:** A. Nemoto, Graduate School of Public Health, Teikyo University, Itabashi-ku, Tokyo, JAPAN|

**CONTROL ID:** 3344537

**TITLE:** Predictive genetic polymorphisms for oxaliplatin treatment efficacy in colorectal cancer – a genome-wide study

**ABSTRACT BODY:**

**Abstract Body:** Oxaliplatin is a platinum drug often given in combination with other anticancer drugs to treat colorectal cancer (CRC). Several candidate gene studies have been conducted to identify susceptibility loci that influence the efficacy of oxaliplatin treatment. However, results have been inconsistent so that predictive genetic markers are currently not available for clinical practice. We conducted a genome-wide association (GWA) analysis to identify novel predictive genetic variants associated with differential prognosis in CRC patients receiving oxaliplatin-based chemotherapy vs. chemotherapy without oxaliplatin.

In total, 1,502 stage II-IV patients that received primary chemotherapy in the German population-based study “DACHS” were included, of whom ~38% received oxaliplatin treatment. A two step-approach was employed for data analysis separately for two subgroups of patients based on stage (1,036 stage II&III and 466 stage IV patients). Firstly multivariable Cox proportional hazards models were used to detect single-nucleotide polymorphisms (SNPs) associated with differential survival according to the type of chemotherapy (oxaliplatin-based vs. others) for three endpoints (overall, CRC-specific, and recurrence-free survival) using an interaction term between SNPs and type of treatment. The False Discovery Rate ( $P < 0.05$ ) was used to account for multiple testing and to select SNPs with significant interaction term for the next step. In the second step, Elastic net regression with a sparsity inducing penalty was applied to the selected SNPs associated with differential treatment outcomes. An internal validation will be conducted using 5-fold cross-validation.

Several SNPs on chromosome 13 q14 showed significant interaction with chemotherapy. One SNP (rs9562572) was selected by the elastic net regression, which showed differential overall survival and CRC specific survival of stage IV patients receiving oxaliplatin-based chemotherapy compared to patients receiving other types of chemotherapy. We did not identify any SNPs among stage II&III patients. Our study suggests that genetic markers might help to identify patients more likely to benefit from oxaliplatin-based chemotherapy. Analyses are ongoing, and findings from the cross-validation will be presented at the conference.

**AUTHORS/INSTITUTIONS:** P. Seibold, L. Jansen, M. Hoffmeister, H. Brenner, J. Chang-Claude, Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, GERMANY|H. Park, Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, GERMANY|D. Edelmann, A. Benner, Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, GERMANY|F. Canzian, Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, GERMANY|M. Schneider, Surgical Oncology Clinic for General, Visceral and Transplantation Surgery, University of Heidelberg, Heidelberg, GERMANY|H. Brenner, German Cancer Consortium (DKTK), Heidelberg, GERMANY|H. Brenner, Division of Preventive Oncology, National Center for Tumor Diseases (NCT), Heidelberg, GERMANY|J. Chang-Claude, Cancer Epidemiology Group, University Cancer Center Hamburg, Hamburg, GERMANY|

**CONTROL ID:** 3344992

**TITLE:** Growth Curve Analysis: Comparing Performances of Mixed-Effect Modelling (MEM) and Structural Equation Modelling (SEM)

**ABSTRACT BODY:**

**Abstract Body:** Background: The last two decades have seen a phenomenal rise in the growth curve models using advanced statistical techniques such as MEM and SEM. There are many parallels between MEM and SEM techniques. However, the proponent of each technique keeps on claiming the superiority of one technique over another. Therefore, keeping the before-mentioned context in mind, the objective of this study is to compare and evaluate the performances of MEM and SEM.

**Methodology**

**Data:** We utilised the secondary data collected for longitudinal neurological progression in HIV-1 and HIV-2 (R01 grants from NIH, USA) patients for the current study. Further, the dataset was simulated for the combinations of the number of participants (30, 50, 100, 200, 500 and 1000), the number of repeated observations (3 to 10) and the magnitude of correlation (0.2, 0.5, 0.8) for repeated measures.

**Analysis and Model Evaluation:** MEM and SEM were analysed using 'lmer' and 'lavaan' package of R-software, respectively. Initially, the live dataset was compared using parameter estimates and their standard errors along with the Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC). Subsequently, the absolute bias, relative bias, absolute inaccuracy index, relative inaccuracy index, and Root Mean Square of Error (RMSE) were used to compare the performance of both the techniques with simulated datasets.

**Result**

**Live dataset:** There were 128 participants at the beginning of the study. However, only 95 subjects provided complete data. The data showed significant and positive correlation (min = 0.47, max = 0.81) for continuous repeated outcome measure. The application of MEM and SEM techniques resulted in the same final models with similar parameter estimates and standard errors.

**Simulated Dataset:** The sample size and statistical technique (MEM versus SEM) contributes significantly to the difference in the bias of parameter estimates. The parameter estimates and standard error were consistent for MEM as compared to SEM for almost all sample size.

**Conclusions:** Both MEM and SEM techniques were found to be efficient in estimating parameter estimates and their standard errors. However, MEM emerged as a favourable technique when the sample size is up to 100 or less in contrast to SEM, where no such trend was observed.

**AUTHORS/INSTITUTIONS:** K. Kishore, Biostatistics, Post Graduate Institute of Medical Education and Research, Chandigarh, Chandigarh, INDIA|P. Marimuthu, K. Thennarasu, D.K. Subbakrishna, Biostatistics, National Institute of Mental Health and Neuro Sciences, Bengaluru, Bengaluru, INDIA|

**CONTROL ID:** 3346502

**TITLE:** Some Weighted Kaplan-Meier Test for Comparing Two Survival Distributions

**ABSTRACT BODY:**

**Abstract Body:** This research presents an approach to improving the weighted Kaplan-Meier test statistics in order to make it a more useful tool for a long-term comparison of two underlying survival distributions in the presence of right-censored data. The procedures are based on the use of some weight function that involves the percentage of censored data as a component. Some versatile procedures for the alternative, not pre-specified, are also discussed. Numerical simulations are conducted to investigate the performance of the proposed procedures. For illustration, the procedures are applied to real-world data in clinical trials, where patients with tongue cancer are divided into two groups according to tumor DNA.

**AUTHORS/INSTITUTIONS:** S. Lee, Mathematics, Illinois Wesleyan University, Bloomington, Illinois, UNITED STATES|

**CONTROL ID:** 3346907

**TITLE:** Bring more data!—a good advice? Removing separation in logistic regression by increasing sample size

**ABSTRACT BODY:**

**Abstract Body:** The parameters of logistic regression models are usually obtained by the method of maximum likelihood (ML). However, in analyses of small data sets or data sets with unbalanced outcomes or exposures, ML parameter estimates may not exist. This situation has been termed 'separation' as the two outcome groups are separated by the values of a covariate or a linear combination of covariates. To overcome the problem of non-existing ML parameter estimates, applying Firth's correction (FC) was proposed. In practice, however, a principal investigator might be advised to 'bring more data' in order to solve a separation issue. We illustrate the problem by means of examples from colorectal cancer screening and ornithology. It is unclear if such an increasing sample size (ISS) strategy that keeps sampling new observations until separation is removed improves estimation compared to applying FC to the original data set. We performed an extensive simulation study where the main focus was to estimate the cost-adjusted relative efficiency of ML combined with ISS compared to FC. FC yielded reasonably small root mean squared errors and proved to be the more efficient estimator. Given our findings, we propose not to adapt the sample size when separation is encountered but to use FC as the default method of analysis whenever the number of observations or outcome events is critically low.

**AUTHORS/INSTITUTIONS:** H. Sinkovec, A. Geroldinger, G. Heinze, Medical University of Vienna, Wien, AUSTRIA|

**CONTROL ID:** 3347275

**TITLE:** Topological data analysis of multi-trial electroencephalographic signals

**ABSTRACT BODY:**

**Abstract Body:**

Topological data analysis (TDA) can decode multi-scale patterns in electroencephalographic (EEG) signals not captured by standard temporal and spectral features. A challenge for applying TDA to groups of long EEG recordings is the ambiguity of performing computationally efficient statistical inference. To address this problem, we advance a unified permutation-based inference framework for testing statistical difference in the TDA feature persistence landscape (PL) of multi-trial EEG signals. The present study applies the topological inference framework to investigate the EEG correlates of speech sensorimotor impairment in post-stroke aphasia patients under a speech altered auditory feedback (AAF) paradigm. Our analysis reveals a significant difference between the PL features extracted from the event-related potential (ERP) response in aphasia vs. control groups over the parietal-occipital and occipital regions when there is no pitch shift in the auditory feedback and over the parietal region when there an upward pitch shift. The findings validate the application of TDA analysis as a robust tool for investigating the neural correlates of speech sensorimotor impairment in neurological patients suffering from speech-language disorders.

**AUTHORS/INSTITUTIONS:** Y. Wang, Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, South Carolina, UNITED STATES|R. Behroozmand, L. Phillip Johnson, J. Fridriksson, Department of Communication Sciences and Disorders, University of South Carolina, Columbia, South Carolina, UNITED STATES|L. Bonilha, Department of Neurology, Medical University of South Carolina, Charleston, South Carolina, UNITED STATES|

**CONTROL ID:** 3347780

**TITLE:** Comparison of scale changes for the simultaneous analysis of qualitative and quantitative variables, using multivariate statistical methods.

**ABSTRACT BODY:**

**Abstract Body:** The management of mixed databases, in other words, mixing of qualitative and quantitative variables regularly presents difficulties for its analysis. One of the alternatives to solve this problem is to homogenize the scale, that is, to change the measurement scale of the variables so that they are all of the same type, this can be achieved, according to the bibliography consulted, by two ways: 1) Categorization and 2) Quantification. The questions of investigation that arose before the alternative of homogenizar the scale, were Which of these two routes presents better results? What advantages and disadvantages does one have over another? What packages are available for these procedures? This paper is the path of a statistical research process that involves the management of mixed multivariate systems, which is a problem that occurs frequently in various areas of knowledge and regularly works improperly. A comparison was made between categorization and quantification as ways to homogenize the scale and try to answer the research questions raised above. The multivariate statistical methods relevant to both cases were identified and exemplified and thus know the advantages and disadvantages of each process. The categorization process was carried out using the quartile method and a multiple correspondence analysis (MCA) was subsequently used, while the HOMALS technique was used for quantification to subsequently use a principal component analysis (PCA for its acronym in English), the analyzes for both processes were performed with the R Project 3.6.1 software. In general, both processes worked properly for this application, although the categorization process is a little longer than the quantification process since the user is assigned the categories. The categorized data had a better representation in the descriptive univariate analysis, while the quantified data had a better representation in the multivariate analysis, which for this particular exercise was paramount. For this reason it is that for this example with real data quantification turns out to be more efficient than categorization.

**AUTHORS/INSTITUTIONS:** E. Morales Garcia, Facultad de Estadística e Informática, Universidad Veracruzana , Xalapa, MEXICO|D. Del Callejo-Canal, M. Canal-Martínez, Instituto de Investigaciones de Estudios Superiores, Económicos y Sociales. Universidad Veracruzana, Xalapa, MEXICO|

**CONTROL ID:** 3348202

**TITLE:** Understanding the role of mediator variables in determining Emergency Department processes

**ABSTRACT BODY:**

**Abstract Body:** Understanding factors that determine wait time in Emergency Department (ED) processes is becoming one major goal of Management in Health Institutions. The main aim of this contribution is to investigate which factors play a relevant role in reducing wait times of people who received services in an ED to better allocate general and specific resources and improve the patients' experience during the stay at the hospital. Every subject who arrives at the hospital ED is registered and every step from the first aid unit is monitored. Time and details of each event related to ED's processes are monitored for a cohort of about 72000 patients during one year. We focus on the time gaps between the arrival of the patient and the treatment time. Along with the times regarding all the process, this dataset includes socio-demographic variables (gender, age, nationality), as well as relevant data regarding the admission in the ED (color code that represent the urgency of the need for assistance, to which hospital unit the patient has been assigned, and whether the patient arrived at the hospital with an ambulance), the diagnosis and the final outcome of the visit. As first, we model wait time data with a survival approach. Preliminary results show that all covariates considered have a significant impact on wait times. In particular, the color code assigned at arrival is a good predictor for wait times, and the most severe situations are treated with more urgency, thus showing the efficacy of patients' initial designation. From a further stratified analysis on different ED areas (Medicine, Surgery, Urgencies, ...) we could better investigate the process, as different wait times behaviors are related to the very different nature of patients' needs. The ED staff shifts are correctly managed to be larger during the most crowded hours, but this is not sufficient to compensate the overloads of the system. We observed a substantial (3.5%) quantity of subject who abandoned the ED before being visited by the clinician. Finally, we focused on ED dropouts to better understand whether these were more related to inefficacy of the ED system rather than to patients' incorrect evaluation of their own status. For this purpose, a decision tree approach was added to identify the "tolerated wait time interval for ED", and other factors like queue length and crowd level.

**AUTHORS/INSTITUTIONS:** A. Nonis, C. Di Serio, Vita-Salute San Raffaele University, Milano, ITALY]

**CONTROL ID:** 3348374

**TITLE:** Modeling the natural history of breast cancer: A study on a cohort of disease-free women in Milan

**ABSTRACT BODY:**

**Abstract Body:** We develop a model for the natural history of breast cancer, where the main events of interest are the start of asymptomatic detectability of the disease and the start of symptomatic detectability. The former kind of detection occurs through screening, while the latter through the emergence of symptoms. We develop a cure rate parametric specification that allows for dependence between the times from birth to the two events, and present some preliminary results of the analysis of data collected as part of a motivating study from Milan. Participants in the study had a varying degree of compliance to a regional breast cancer screening program. The subjects' ten-year trajectories have been obtained from administrative data collection performed by the Italian national health care system. We focus on the outcome of the mammograms, and after taking into account the selection process into the study we develop the likelihood contributions of the possible observed trajectories to perform maximum likelihood inference on the underlying process. The estimated parameters of the underlying disease process allow for the study of the effect of different examination schedules on a population of asymptomatic subjects, starting from different ages.

**AUTHORS/INSTITUTIONS:** M. Bonetti, Dondena Research Center, Milan, MI, ITALY|M. Bonetti, D. Grigorova, Social and Political Sciences, Bocconi University, Milan, MI, ITALY|M. Bonetti, Bocconi Institute for Data Science and Analytics, Milan, MI, ITALY|L. Bondi, Decision Sciences, Bocconi University, Milan, MI, ITALY|A.G. Russo, Epidemiology, ATS MILANO - Città metropolitana, Milan, MI, ITALY|

**CONTROL ID:** 3348936

**TITLE:** Statistical Models To Identify Factors Affecting The Success Of Dental Implants

**ABSTRACT BODY:**

**Abstract Body:** A dental implant is a surgical component made of titanium and other materials that is positioned into the jawbone beneath the gum line to support a dental prosthesis. In the first step the implant is anchored into the jawbone and three to six months later, after the implant has fully fused with the bone, the dental prosthesis is attached to a metal post protruding from the implant. A common approach to replace not only one tooth by an implant is the so called All-on-4<sup>®</sup> treatment concept. This concept provides patients with a replacement of an entire row of teeth based on four implants. A further benefit for patients is, that implants are considered long-term replacements that can last up to a lifetime.

The success of the implantation process can be measured based on the percentage of the occurrence of peri-implantitis (a destructive inflammation affecting the hard and soft tissue surrounding the implants), peri-mucositis (a reversible inflammation affecting only the soft tissue surrounding the implants) and survival of the implant. Additionally, the marginal bone level change can be used as a continuous outcome parameter. All of these outcome parameters are usually evaluated as part of an annual dental check-up.

From a statistical point of view a model for identifying factors affecting implant success often violates the assumption of independency (i.e. more than one implant per patient). The idea of using a simple aggregation approach based on means or sums to account for the situation of multiple implants per patient leads to a loss of power and information. Hence the model has to account for this situation, so mixed models or generalized estimating equation techniques should be used. Additionally, to the violation of the assumption of independency often the effect over time should be estimated and included in the statistical model.

Based on the data of a 5-year prospective study various statistical models are presented and compared regarding their goodness of fit.

**References:**

Cho Y., Kim: HY.: Analysis of periodontal data using mixed effects models, J. Periodontal Implant Sci. 2015; 45, pp. 2-7.

Krennmair S, Hunger S, Forstner T, Malek M, Krennmair G, Stimmelmayr M.: Implant health and factors affecting peri-implant marginal bone alteration for implants placed in staged maxillary sinus augmentation: A 5-year prospective study, Clin. Implant. Dent. Relat Res., 2019,21(1), pp. 32-41.

**AUTHORS/INSTITUTIONS:** T. Forstner, Department of Applied Systems Research and Statistics, Johannes Kepler University, Linz, AUSTRIA|S. Krennmair, Ludwig-Maximilian University (LMU), Munich, GERMANY|

**CONTROL ID:** 3349011

**TITLE:** GENOME-BASED ALGORITHM FOR GENE SELECTION AND CLASSIFICATION OF COLORECTAL CARCINOMA

**ABSTRACT BODY:**

**Abstract Body:** This study presents a method for optimal selection of gene subsets to enhance the non-clinical diagnostic classification of colorectal cancer using gene expression level of 40 tumour and 22 normal colon tissues for 2,000 gene expression profiles obtained with an Affymetrix oligonucleotide array. An efficient Hybrid multi-objective Support vector Machine (SVM) feature selection and classification algorithm was employed to determine the most informative few genes subsets that are highly relevant to the 62 (tumour or normal) responses of the gene expression levels. The genes selection was done in two stages with the first stage using the Bayesian T-test to prune the non-informative genes and the second stage employed the multi-objective optimization that allows sequential addition of genes for optimal determination of the pre-selected gene subsets. In a Monte-Carlo experiment, a pre-selection of the features was performed with the filter method based on Sidak alpha value to reduce the number of false-positive features in the data. The optimal values of tuning parameters for both the SVM cost and Radial Basis Function (RBF) kernel were determined by grid search in 10-fold cross-validation. The SVM with RBF kernel was then fitted sequentially to select the set of near-optimal genes that are correlated with the response class. The optimally selected genes yielded an accuracy of 90.1% on the test data that were never used in the building process of the algorithm. Also, an estimated average of 86.94%, 91.92%, 85.87%, 92.64% and 91.56% was obtained for Sensitivity, Specificity, Positive predictive value, negative predictive value and Cross-validated Area under the curve (CVAUC) respectively. Furthermore, the results obtained from the principal component analysis and the complete linkage hierarchical clustering indicated perfect discrimination of the two clinical response groups of the colorectal cancer status of the patients. The sets of optimally selected gene subsets in the data can be further investigated by molecular biologist to establish the pathology of these genes with respect to their respective tumour classes.

**AUTHORS/INSTITUTIONS:** A.W. Banjoko, Statistics, University of Ilorin, Ilorin, Kwara, NIGERIA|

**CONTROL ID:** 3349076

**TITLE:** The transmuted alpha power-G family of distributions

**ABSTRACT BODY:**

**Abstract Body:** This article proposes a new class of models called the transmuted alpha power-G (TAPO-G) family of distribution for modeling lifetime processes. This class of model extends the well-known existing distributions. A comprehensive statistical structural property of the proposed model is established. The parameters of the proposed model are obtained by maximum likelihood method. The performance of the maximum likelihood estimators was accessed by mean squared errors and biases of the Monte Carlo simulation study. The proposed model usefulness was examined by real life data.

**AUTHORS/INSTITUTIONS:** J.T. Eghwerido, E. Efe-Eyefia, S.C. Zelibe, Mathematics and Computer Science, Federal University of Petroleum Resources, Effurun, Delta, NIGERIA]

**CONTROL ID:** 3349157

**TITLE:** The Change in Longitudinal CD4 Cell Count and Time to Death among HIV Infected Children Initiating ART

**ABSTRACT BODY:**

**Abstract Body:** The main aim of this study was to identify the determinants of the change in longitudinal CD4 cell count and survival time of HIV-infected children under ART. A cohort of 201 HIV infected children aged less than 15 years were followed from October 2013 to March 2017 at Adama Hospital in Ethiopia. The study design was retrospective. Separate linear mixed effect model were done for the longitudinal outcome of the CD4 cell count. Cox PH model was conducted for the time-to-event outcome. Joint modeling was performed both for the longitudinal and survival outcomes and results were compared with the separate analysis. However, with the specific interest in identifying the determinants and characterizing the relationship between a longitudinal CD4 cell count and time-to-event outcomes, the study focused on joint modelling. The finding from the joint modeling indicated the estimated association parameter is -0.10 and this was statistically significant. This shows higher values of CD4 cell count associated with less risk of death. Seemingly we found that; observation time, age at HIV/AIDS diagnosis, WHO clinical stages, history of TB and functional status were determinants for the mean change in the square root of CD4 cell count. Furthermore WHO clinical stages, functional status, history of TB and BMI have significant negative impact on the survival probabilities of HIV infected children.

**KEYWORDS:** HIV/AIDS, children, joint model, CD4 cell count, time to death

**AUTHORS/INSTITUTIONS:** D.A. Teni, B. Belete, Statistics, Arba Minch University, Arba Minch, SNNPR, ETHIOPIA|T. Nigusie, Statistics, Bule Hora University, Oromiya, ETHIOPIA|

**CONTROL ID:** 3349642

**TITLE:** Identify the best genotype via genomic prediction

**ABSTRACT BODY:**

**Abstract Body:** An iterative strategy using genomic prediction is proposed to identify the best genotype from a candidate population, which is derived from the expected improvement (EI) criterion of Bayesian optimization. The iteration search consists of two steps. A genomic BLUP (GBLUP) prediction model is first built using the phenotype and genotype data of a training population. An EI criterion is then employed to select the individuals, which are expected to include the best genotype. If the best genotype is not identified at the current step, the selected individuals are phenotyped and added with the training population to update the GBLUP prediction model. In this study, we propose a forward EI criterion based on the distribution of predicted genotypic values. Our proposed forward EI criterion is shown to be advantageous over some other existing criteria, mainly because that it considers the correlation among the candidate individuals, and is not affected by the random or environmental variation. Two real datasets of rice and wheat are analyzed to illustrate our proposed procedure.

**AUTHORS/INSTITUTIONS:** C. Shen, C. Liao, Department of Agronomy, National Taiwan University, Taipei, TAIWAN|

**CONTROL ID:** 3349854

**TITLE:** A flexible modelling framework using linear splines for censored regression models to estimate serological profiles.

**ABSTRACT BODY:**

**Abstract Body:** Background. In 2011, Malawi introduced the 13-valent pneumococcal conjugate vaccine. We have completed a serosurvey, measuring serotype-specific and anti-protein antibody levels (IgG) to assess study vaccine response, subsequent waning and naturally acquired immunity.

Methods. Blood samples from children were collected during an 18-month period in 2017/18 in Blantyre, Malawi by randomly sampling children of various ages between 0 and 5 years. A total of 556 samples were collected and are included in this analysis. Capsule-specific IgG to all 13 vaccine-serotype (VT) *Streptococcus pneumoniae* serotypes (1, 3, 4, 5, 6A, 6B, 7F, 9V, 14, 18C, 19A, 19F, 23F) as well as two common non-vaccine serotypes (NVT; 12F, 33F) were measured using standardised enzyme immunoassay and direct binding electrochemiluminescence-based multiplex assay methods.

We developed a modelling framework to estimate population-averaged serological profiles. To account for the presence of a lower limit of detection (DL), we used censored regression models. To estimate the number and location of changepoints in the serological profiles, we used linear splines to regress IgG on age, treating the knots as model parameters estimated using maximum likelihood. Confidence intervals for the changepoints were derived using the bootstrap percentile method. To optimise fit while penalising model complexity, Akaike information criterion was used to select the number of knots (0, 1, 2 or 3 knots).

Results. Our models recapitulate the observed geometric mean antibody profiles. Importantly, we can quantify the number and location of changepoints and the rates of IgG increase and waning. By computing confidence intervals for all model parameters, our approach allows principled statistical inference.

With one exception, all VT serotypes show a clear vaccine-induced response. We observe several distinct profiles for VT serotypes, corresponding to potentially important biological differences. IgG concentrations for both NVT serotypes were too low for reliable modelling.

Conclusion. We present a flexible modelling framework to estimate both the number and location of changepoints in serological profiles. Key benefits of our approach are the principled statistical analysis of serological profile data which are increasingly common and the grouping of profiles, giving insights into the underlying biology.

**AUTHORS/INSTITUTIONS:** M.Y. Henrion, T.D. Swarthout, S.N. Barnaba, J.E. Meiring, M.A. Gordon, Malawi - Liverpool - Wellcome Trust Clinical Research Programme, Blantyre, MALAWI|M.Y. Henrion, Liverpool School of Tropical Medicine, Liverpool, UNITED KINGDOM|T.D. Swarthout, R.S. Heyderman, NIHR Global Health Research Unit on Mucosal Pathogens, University College London, London, UNITED KINGDOM|S.N. Barnaba, Chancellor College, University of Malawi, Zomba, MALAWI|D. Goldblatt, Great Ormond Street Institute of Child Health Biomedical Research Centre, University College London, London, UNITED KINGDOM|D. Thindwa, Centre for Mathematical Modelling of Infectious Diseases, Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UNITED KINGDOM|J.E. Meiring, Oxford Vaccine Group, Department of Paediatrics, University of Oxford, Oxford, UNITED KINGDOM|M.A. Gordon, N. French, Institute of Infection and Global Health, University of Liverpool, Liverpool, UNITED KINGDOM|

**CONTROL ID:** 3350124

**TITLE:** A Bayesian nonparametric approach for high-dimensional causal mediation

**ABSTRACT BODY:**

**Abstract Body:** We propose a Bayesian nonparametric approach to the problem of high-dimensional causal mediation. The model for the observed data is an extension of the enriched Dirichlet process (EDP) mixtures introduced in Wade et al. (2011). In the context of these EDP mixtures, we propose an approach for selection of mediators and introduce a set of assumptions sufficient to identify both joint (all mediators) and individual (one mediator) natural direct and indirect effects. The approach is evaluated via simulations.

**AUTHORS/INSTITUTIONS:** S. Roy, M. Daniels, Statistics, University of Florida, Gainesville, Florida, UNITED STATES|J. Roy, Biostatistics & Epidemiology, Rutgers University, Piscataway, New Jersey, UNITED STATES|

**CONTROL ID:** 3351772

**TITLE:** Long-term frailty modeling using a non-proportional hazards model: Application with a melanoma dataset

**ABSTRACT BODY:**

**Abstract Body:** The semiparametric Cox regression model is often fitted in the modeling of survival data. One of its main advantages is the ease of interpretation, as long as the hazards rates for two individuals do not vary over time. In practice the proportionality assumption of the hazards may not be true in some situations. In addition, in several survival data is common a proportion of units not susceptible to the event of interest, even if, accompanied by a sufficiently large time, which is so-called immune, "cured," or not susceptible to the event of interest. In this context, several cure rate models are available to deal with in the long term. Here, we consider the generalized time-dependent logistic (GTDL) model with a power variance function (PVF) frailty term introduced in the hazard function to control for unobservable heterogeneity in patient populations. It allows for non-proportional hazards, as well as survival data with long-term survivors. Parameter estimation was performed using the maximum likelihood method, and Monte Carlo simulation was conducted to evaluate the performance of the models. Its practice relevance is illustrated in a real medical dataset from a population-based study of incident cases of melanoma diagnosed in the state of São Paulo, Brazil.

**AUTHORS/INSTITUTIONS:** V. Calsavara, E. Bertolli, Department of Epidemiology and Statistics, A.C.CAMARGO CANCER CENTER, São Paulo, BRAZIL|V.L. Tomazella, Estatística, Universidade Federal de São Carlos, São Carlos, São Paulo, BRAZIL|E. Milani, Universidade Federal de Goiás, Goiânia, BRAZIL|

**CONTROL ID:** 3356000

**TITLE:** Improved design approach for field experiments conducted to assess cultivar tolerance and resistance to root lesion nematode (*Pratylenchus* spp.)

**ABSTRACT BODY:**

**Abstract Body:** The root-lesion nematodes (RLN) *Pratylenchus thornei* and *P. neglectus* are widely distributed within cropping regions of Australia and limit grain production. Field experiments comparing the performance of cultivars in the presence of RLN investigate management options for growers by identifying cultivars with resistance (limiting nematode reproduction) and tolerance (yielding well in the presence of nematodes). Conventional approaches used in these experiments can result in biased estimates or the confounding of cultivar yield potential, tolerance and resistance. We propose a novel experimental design approach for RLN trials in which the observed RLN density, estimated at sowing, is used to condition the randomization of cultivars to field plots. The approach ensures that cultivars are exposed to an approximately equivalent range of RLN burdens which allows valid assessments of relative cultivar performance to be derived. Data from a field experiment designed using the proposed approach was analysed in a linear mixed model framework. The design approach allowed relative resistance to be modelled over a consistent range of pre-sowing RLN densities, so that groups of cultivars sharing similar resistance levels could be identified. A comparison of slopes of yield response curves of cultivars belonging to the same resistance class identified differing tolerance levels for cultivars with equivalent exposures to both pre-sowing and post-harvest RLN densities.

**AUTHORS/INSTITUTIONS:** K. Reeves, Curtin University, Bentley, Western Australia, AUSTRALIA|K. Reeves, R. Loughman, Department of Primary Industries and Regional Development, South Perth, Western Australia, AUSTRALIA|C. Forknall, A. Kelly, Queensland Department of Agriculture and Fisheries, Toowoomba, Queensland, AUSTRALIA|K. Owen, Centre for Crop Health, University of Southern Queensland, Toowoomba, Queensland, AUSTRALIA|J. Fanning, G. Hollaway, Agriculture Victoria, Horsham, Victoria, AUSTRALIA|

**CONTROL ID:** 3356178

**TITLE:** Bayesian longitudinal models for assessing the artisanal and industrial sardine fishing in the Mediterranean Sea

**ABSTRACT BODY:**

**Abstract Body:** Small pelagic fish species have been proven to be key elements of the Mediterranean pelagic ecosystem due to their high bulk of biomass at the mid-trophic level, which provides an important energy connection between the lower and upper trophic levels. Overfishing is as a key factor in the collapse of several populations of small pelagics, often in combination with environmental fluctuations and the global climate change process. In the Mediterranean Sea, catches are dominated by small pelagics, representing nearly the 49% of the total harvest. Among them, the European sardine (*Sardina pilchardus*) is one of the most commercial species showing high over-exploitation rates over the last years. Mediterranean fisheries are highly diverse and geographically varied, not only because of the existence of different marine environments, but also because of different socio-economic situations and fishery status.

Within this context, we analyse European sardine landings from 1970 to 2014 by countries in the Mediterranean Sea, from both the artisanal and the industrial fisheries. The statistical analysis is based on the Bayesian linear mixed-effects model and the shared-parameter model framework to joint modelling longitudinal data. This approach builds a relationship of association between the longitudinal process of the industrial fisheries and that of the artisanal fisheries, through the random effects.

We compare two different models, with and without serial correlation (i.e., a joint mixed linear model and an autoregressive joint longitudinal model), by means of the posterior distribution of the conditional residuals and the Bayes factors. The computation of the Bayes factors requires the non-trivial numerical estimation of the marginal likelihoods. Finally, we discuss a few methods for model comparison for this class of models.

**AUTHORS/INSTITUTIONS:** G. Calvo, C. Armero, Universitat de València, Valencia, Valencia, SPAIN|M. Pennino, Instituto Español de Oceanografía, Vigo, SPAIN|L. Spezia, Biomathematics & Statistics Scotland, Scotland, Aberdeen, UNITED KINGDOM|

**CONTROL ID:** 3356241

**TITLE:** A new workflow combining R packages for statistical analysis of metabolites

**ABSTRACT BODY:**

**Abstract Body:** In metabolomics, the establishment of a relationship between the metabolome and human basic traits (such as age or sex, or cultivar in foods) is a key research question.

For this, we present a new statistical workflow, using the CRAN packages survival, tram, mlt, and multcomp. The workflow is demonstrated using a large-scale study with public data (Thévenot et al.) and using the KarMeN cross-sectional data (Frommherz et al.). Specifically, the association between the covariate age for the KarMeN sample and two selected metabolites (without and with detection limit) is presented.

The workflow makes no a priori assumptions about the distribution of the metabolites; since it is unrealistic to assume the same error distribution for each of them. Instead, a metabolite-specific data-driven transformation function of each metabolite will be involved by the so-called most likely transformation (with the help of the packages tram and mlt, compare Hothorn et al.).

The workflow takes into account the fact that the metabolites are a mixture of completely measured metabolites, left-censored metabolites (with values below the limit of detection or quantification), and metabolites with many ties (with the help of the package survival).

The association between the variable of interest and the multiple metabolites is established simultaneously. The adjustment for multiple comparisons considers another intrinsic property of the metabolites, namely that they are often correlated in subgroups. This information has been included and leads to adjustments for multiplicity that can be less conservative compared to approaches that ignore it (with the help of the package multcomp).

**References:**

Frommherz L. et al. (2016) Plos One 11(4):e0153959

Thévenot et al. (2015) J Proteome Res, 14(8):3322-3335

Hothorn T et al. (2018) Scand J Stat 45(1):110-134

**AUTHORS/INSTITUTIONS:** P. Ferrario, Department of Physiology and Biochemistry of Nutrition, Max Rubner-Institut, Karlsruhe, GERMANY|

**CONTROL ID:** 3356244

**TITLE:** Subgroup Identification Method Based on Multiple Adaptive Regression Spline

**ABSTRACT BODY:**

**Abstract Body:** Objective To propose a subgroup identification method based on multivariate adaptive regression spline model for survival data. Methods The virtual twins method was used to estimate the subgroups of patients. Multivariate adaptive regression spline model is applied to do the subgroup identification, with the purpose of screening the Subgroup related covariates and predicting the subgroup information of patient. Finally, we established a Cox model based on the obtained subgroup information, and evaluated the benefit level of each subgroup by HR value, and conducted real data analysis. Results The simulation results showed that the method performed well in terms of Power, subgroup related covariate recognition rate, patient subgroup correct rate and type I error control. Conclusion This method can effectively and reliably evaluate subgroup benefit levels, identify subgroup related covariates, and recommend appropriate treatments for patients. It can provide an effective decision-making basis for precision medicine.

**AUTHORS/INSTITUTIONS:** S. Zhou, Southern Medical University, Guangzhou, CHINA|

**CONTROL ID:** 3356277

**TITLE:** KNOWLEDGE ON HEPATITIS B INFECTION AND VACCINATION UPTAKE AMONG HEALTHCARE WORKERS AT KILIMANJARO CHRISTIAN MEDICAL CENTRE MOSHI, TANZANIA.

**ABSTRACT BODY:**

**Abstract Body:** Background: Hepatitis B infection is a global problem exposing more than two billion people worldwide, with the highest prevalence in Asia and Sub Saharan Africa. Hepatitis B virus infection is transmitted through infected blood and bodily fluids. Health care workers are at higher risk to hepatitis B virus infection due to exposure to blood and bodily fluids from patients they attend daily. We aimed to assess knowledge on hepatitis B infection and vaccination uptake among healthcare workers at Kilimanjaro Christian Medical Centre, Moshi Tanzania. Methodology: A cross sectional study which was conducted using a standardized structured self-administered questionnaire with 16 questions on knowledge each carried one mark. The total score was converted to percentage and grouped as poor ( fair (50-74%) and good ( $\geq 75\%$ ) as was adapted from article published by Debes et al,2016 Data obtained from the study was analyzed by SPSS version 20.

Results: This study included a total of 442 healthcare workers at Kilimanjaro Christian Medical Centre referral and teaching hospital. The median age of participants was 37 and interquartile range (IQR) of 31-46 years. Over 60% of all participants were females and above three quarter (78.5%) had tertiary education (College/ University) level. 109(24.7%) of the respondents had poor knowledge on hepatitis B infection, 219(49.6%) had fair knowledge and 114(25.8%) had good knowledge. Out of 295 participants who ever received hepatitis B vaccination, 70.5% received three shots, 19.5% received two shots and 10.0% received one shot.

Conclusion: Only 25.8% had good knowledge and there was poor vaccination uptake among health care workers in Kilimanjaro Christian Medical Centre, Moshi Tanzania. More efforts should be devoted to improve knowledge on hepatitis B infection by doing continuous medical education. Furthermore hepatitis B vaccination should be mandatory to all healthcare workers to prevent them from contracting hepatitis B infection at their working environment.

**AUTHORS/INSTITUTIONS:** E.M. Temu, internal medicine, kilimanjaro christian medical university college, Moshi-Kilimanjaro, TANZANIA, UNITED REPUBLIC OF|

**CONTROL ID:** 3356283

**TITLE:** Case-deletion diagnostics for mixed-effects location scale models

**ABSTRACT BODY:**

**Abstract Body:** Mixed-effects location scale models allow simultaneous modelling of between-subject and within-subject variability. These models include log-linear models for the between-subject and within-subject variability. The log-linear models could potentially include covariates. We explore Cook-type influence diagnostics for the mixed-effects location scale model. We also extend these diagnostics to the multivariate longitudinal data case of the mixed-effects location scale model. A real dataset is analyzed to illustrate the influence diagnostics.

**AUTHORS/INSTITUTIONS:** F.N. Gumedze, Statistical Sciences, University of Cape Town, Rondebosch, Western Cape, SOUTH AFRICA

**CONTROL ID:** 3356345

**TITLE:** Polygenic risk scores – Is there a need for a more accurate classification within ethnicities?

**ABSTRACT BODY:**

**Abstract Body:** Polygenic risk scores for complex diseases are widely used in preclinical and clinical research to stratify individuals according to their genetic risk for targeted prevention, therapy, or prognosis. However, they are usually derived and validated within a specific ethnic background, and translation into other ethnicities has been shown to be problematic. Furthermore, even the transfer between populations in the same country can be challenging, as shown, for instance, for Finland (1) and Great Britain (2).

According to former studies, at least slight genetic differences are present between different parts of Germany (3). However, the implications for polygenic risk scores have not been evaluated so far. Therefore, this study aims at investigating the impact of geographic regions within Germany on the distribution of polygenic risk scores for common complex diseases.

The German Neonatal Network examines the development of very low birth weight infants with 64 study centers spread across Germany. Umbilical cord tissue frozen after birth is used to genotype the DNA of the infants. Affymetrix Axiom<sup>TM</sup> Genome-Wide CEU 1 Array Plate 2.0 and Illumina Infinium® Global Screening Array-24 v1.0/v2.0 were used for chip genotyping.

The continuously growing database already contains genetic data of 10,259 (51,2%) from 20,000 included very low birth weight infants. Within this database, we construct polygenic risk scores for common complex diseases, based on the GWAS Catalog (4), and compare their distributions between various areas within Germany. Results will provide insight into the transferability of polygenic risk scores between populations but also into the genetic architecture of the investigated traits.

1. Kerminen S, Martin AR et al. Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. *Am J Hum Genet* 2019; 104(6):1169-81.
2. Haworth S, Mitchell R et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat Commun* 2019; 10:333.
3. Steffens M, Lamina C et al. SNP-Based Analysis of Genetic Substructure in the German Population. *Hum Hered* 2006; 62(1):20-9.
4. Buniello A, MacArthur JAL et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*, 2019, 47:D1005-D1012.

**AUTHORS/INSTITUTIONS:** T.K. Rausch, I.R. König, Institut für medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, GERMANY|T.K. Rausch, W. Göpel, Klinik für Kinder- und Jugendmedizin, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, GERMANY|

**CONTROL ID:** 3356457

**TITLE:** Analytical methods used in handling missing data in estimating the prevalence of HIV/AIDS for demographic and cross-sectional surveys: A systematic review

**ABSTRACT BODY:**

**Abstract Body:** Background: Surveys studies often have a problem of missing data. Few studies report the proportion of missing data and even fewer describe the methods used to adjust the results for missing data. The objective of this review was to determine the analytical methods used in handling missing data in surveys for estimating the prevalence of HIV/AIDS in the population.

Methods: We searched for population, demographic and cross-sectional surveys published from January 2000 to April 2018 in Pub Med/Medline, Web of Science core collection, Latin American and Caribbean Sciences Literature, Africa-Wide Information and Scopus, and by reviewing references of included articles. All potential abstracts were imported into Covidence and screened by two independent reviewers using pre-specified criteria. Disagreements were resolved through discussion. A piloted data extraction tool was used to extract data and assess risk of bias. Data were analysed through a quantitative approach; variables were presented and summarised using figures and tables.

Results: A total of 3426 citations were identified, 194 duplicates removed, 3232 screened and 69 full articles were obtained. Twenty-four studies were included. The response rate for an HIV test of the included studies ranged from 32% to 96% with the major reason for the missing data being consent refusal to an HIV test. Complete case analysis was the primary method of analysis used, multiple imputations 11(46%) was the most advanced method used followed by the Heckman's selection model 9(38%). Single Imputation and Instrumental variables method were used in only two studies each, with 13(54%) other different methods used in several studies. Forty-two percent of the studies applied more than two methods in the analysis, with a maximum of 4 methods per study. Only 6(25%) studies conducted a sensitivity analysis, while 11(46%) studies had a significant change of estimates after adjusting for missing data.

Conclusion: Missing data in survey studies is still a problem in disease estimation. Our review outlined a number of methods that can be used to adjust for missing data, however, more information and awareness are needed to allow informed choices on which method to be applied for the estimates to be more reliable and representative.

**AUTHORS/INSTITUTIONS:** N.R. Mosha, O. Aluko, R. Machezano, T. Young, Epidemiology and Biostatistics, Stellenbosch University, Capetown, SOUTH AFRICA|N.R. Mosha, J. Todd, National Institute for Medical Research, Mwanza, TANZANIA, UNITED REPUBLIC OF|J. Todd, London School of Hygiene and Tropical Medicine, London, UNITED KINGDOM|

**CONTROL ID:** 3356708

**TITLE:** Handling Coarsened Age Information in the Analysis of Emergency Department Presentations

**ABSTRACT BODY:**

**Abstract Body:** Administrative databases offer vast amounts of data that provide opportunities for cost-effective insights. They simultaneously pose significant challenges to statistical analysis such as the redaction of data because of privacy policies and the provision of data that may not be at the level of detail required. For example, ages in years rather than birthdates available at event dates can pose challenges to the analysis of recurrent event data. Hu and Rosychuk (2016) provided a strategy for estimating age-varying effects in a marginal regression analysis of recurrent event times when birthdates are missing. Their research was motivated by emergency department (ED) visits made by children and youth and privacy rules that prevented birthdates to be released. With recent changes in the data access rules for Alberta data, we are able to request a new extract of the mental health ED data that includes patient birthdates for April 2010 to March 2017. This new extract is employed to evaluate the estimates using the Hu and Rosychuk approach for coarsened ages with estimates under the true, known ages. We find that overall the Hu and Rosychuk approach for coarsened ages performed well and captured the key features of the relationships between ED visit frequency and covariates.

**AUTHORS/INSTITUTIONS:** A. Chen, Pediatrics, University of Alberta, Edmonton, Alberta, CANADA|J.W. Bachman, Individual, Edmonton, Alberta, CANADA|X. Hu, Simon Fraser University, Burnaby, British Columbia, CANADA|R.J. Rosychuk, University of Alberta, Edmonton, Alberta, CANADA|

**CONTROL ID:** 3356859

**TITLE:** Order statistics approach to modeling and prediction of early mood swing

**ABSTRACT BODY:**

**Abstract Body:** Owing to the increase rate of attempted and real suicide in Nigeria, a situation many attributed to the harsh economic condition in the country; then it is obvious that the management of extreme happiness and sadness has become a problem that needs to be checked. Thus based on the tail flexibility property of exponential power distribution (EPD), this paper obtain the order statistics survival model of mood swing and established its behavioral pattern for hypothetical and real life cases; while controlling for clients (patients) response bias aimed at deliberately misleading the psychiatrist by giving false claim via conditional order statistics tool. The study modeled and predict early mood swing to facilitate quick and early medical intervention before it builds up to something more difficult to control. The obtained model via order statistics approach which showed a square wave form for the hypothetical case and an approximate monotone damping model for the real life data is more representative of the real life situation compared to existing cubic spline model

**AUTHORS/INSTITUTIONS:** T.A. Soyinka, F. Apantaku, A. Akintunde, O.A. Wale-Orojo, K. Yusuff, O. Asiribo, Department of Statistics, Federal University of Agriculture, Abeokuta, Ogun state. Nigeria., Abeokuta, Ogun, NIGERIA|A.A. Olosunde, O. Obilade, Department of Mathematics, Obafemi Awolowo University, Ile-ife, Osun, NIGERIA|

**CONTROL ID:** 3357612

**TITLE:** Statistical interpretations of mortality rate for treatment resistant diseases with real world data and associated biomarkers research

**ABSTRACT BODY:**

**Abstract Body:** Treatment resistance is a challenging issue encountered by physicians, patients, and researchers in many highly disabling, serious conditions with significant morbidity and mortality associations. Here we contrast two widely discussed treatment resistant diseases: Major depression disorder (MDD) in psychiatry and ovarian cancer (OC) in oncology. Real world data (RWD) plays an increasingly important role in the regulatory approval for marketing applications of new therapies in recent years, especially for rare diseases, among which treatment-resistant-diseases are a major category. "Treatment Resistant Depression" (TRD) is a severe subset of the MDD. Several cohort studies have been published regarding the economic burden and mortality rates in TRD versus non-TRD (NTRD) patients. In oncology, approximately 25% of early stage ovarian cancer patients and more than 80% of advanced ovarian cancer patients experience disease relapse within two years of the initial treatment of chemotherapy with carboplatin and paclitaxel. Not surprising, it is often claimed that patients with treatment-resistant disease, whether it is depression or cancer, have shorter life-span than patients with treatment-sensitive disease. However, the measure and comparison of the mortality rates for disease resistant versus disease sensitive patients with real world data in these two cases (MDD and OC) are quite different. The statistical issues that have not been much discussed in the literature. In this talk, we will delineate some of these challenges and provide some recommendations related to the treatment-resistance etiology research in biomarkers.

**AUTHORS/INSTITUTIONS:** L. Peng, Biostatistics, Eisai, Inc., Woodcliff Lake, New Jersey, UNITED STATES|

**CONTROL ID:** 3357630

**TITLE:** Fast Estimation and Bootstrap-based Confidence Intervals for Threshold Linear Regression Models with Higher-order Terms

**ABSTRACT BODY:**

**Abstract Body:** Threshold regression models are useful for describing a nonlinear relationship between an outcome and a predictor, especially when it is of interest to study a threshold parameter. Previous research on threshold regression models has mostly focused on models that are linear before and after the threshold. In this work we study threshold linear regression models that allow higher order trends both before and after the threshold. Estimation of threshold regression models is a non-convex and non-smooth optimization problem. Finding the true maximum likelihood estimator requires a grid search, which is very slow if done in a brute force way. Following our previous works on threshold model estimation, we develop fast grid search algorithms for threshold linear regression models with higher order trends and assess their performance. We further develop model-robust confidence intervals for model parameters, as well as pointwise and simultaneous confidence bands for mean functions using Efron's bootstrap technique. We conduct Monte Carlo studies to demonstrate the performance of the proposed methods, and illustrate the application of the models using a real dataset from environmental physics.

**AUTHORS/INSTITUTIONS:** H. Son, Department of Biostatistics, University of Washington, Seattle, Washington, UNITED STATES|Y. Fong, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, UNITED STATES|

**CONTROL ID:** 3357921

**TITLE:** Geographically Weighted Regression Approach to the Analysis of Treatment Effects in On-farm Experiments

**ABSTRACT BODY:**

**Abstract Body:** We present a unifying approach to inference for the analysis of on-farm strip experiments, adapting the core ideas of local likelihood or geographically weighted regression. We propose a statistical model that allows spatial nonstationarity in modelled relationships and estimates spatially-varying parameters governing these relationships. A crucial step is bandwidth selection in implementing these models, and we develop bandwidth selection methods for scenarios relevant to modelling yield monitor data in on-farm experiments. Local t-scores have been introduced for inferential purpose and the associated problem of multiple testing has been described in the context of analyzing on-farm experiments. We have shown in this paper how local p-values can be adjusted to overcome this problem. To illustrate the applicability of our proposed method, we have analyzed two on-farm datasets. Graphical displays are created to guide practitioners to make informed decisions on optimal management practices.

**AUTHORS/INSTITUTIONS:** S. rakshit, SAGI West, School of Molecular and Life Sciences, Curtin University, Bentley, Western Australia, AUSTRALIA]

**CONTROL ID:** 3357934

**TITLE:** Modelling Soybean Growth: A Nonlinear Mixed Model Approach

**ABSTRACT BODY:**

**Abstract Body:** Field experiments of soybean were conducted in Arid Land Research Center, Tottori, Japan, under two experimental conditions (drought and control). The growth was monitored by an unmanned aerial vehicle (UAV), or drone, measuring each day the plant height and canopy area of 201 soybean varieties for which whole-genome sequence data are also available.

In this work, we propose to model the plant height as a nonlinear function of time and environmental variables such as daily temperature, solar radiation, soil moisture. This type of modelling includes mechanistic parameters having a concrete meaning in relation to the growth of the plant. We combine it with random effects to describe inter-plant variability.

The objective is to quantify the respective roles of the environmental variables in the plant development. Parameter estimation in nonlinear mixed models is however not straightforward, especially due to the model nonlinearity and to the random effects. SAEM algorithm (Stochastic Approximation EM) is widely used in this context and generally gives interesting results. In our approach, one more difficulty is that plant height is observed at 26 different dates creating time dependence across observations. We propose a specific version of SAEM that takes into account this form of time series dependence.

Performances of the proposed algorithm are first assessed on simulated data and preliminary results on soybean data are presented.

**AUTHORS/INSTITUTIONS:** M. Delattre, J. Tressou, MIA Paris, AgroParisTech, INRAE, Université Paris-Saclay, Paris, FRANCE|H. Iwata, J. Tressou, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, JAPAN|

**CONTROL ID:** 3357974

**TITLE:** Exploring the spatial patterning in racial differences in cardiovascular health between blacks and whites across the United States: The Reasons for Geographic and Racial Differences in Stroke Study

**ABSTRACT BODY:**

**Abstract Body:** Background and Purpose:

Cardiovascular health (CVH) disparities between blacks and whites have persisted in the United States (US) for some time, and although there have been remarkable improvements in addressing cardiovascular disease, it still remains the leading cause of death in the US. Additionally, well-documented disparities are unfortunately widening incidence gaps across certain regions of the US. Our focus was on answering the following questions: (1) How much spatial heterogeneity exists in the racial differences in CVH between blacks and whites across this country? and (2) Is the spatial heterogeneity in the racial differences significantly explained by living in the Stroke Belt?.

Methods and Results:

To explore the spatial patterning in the racial differences in CVH between blacks and whites across the country, we utilized a national, population-based, longitudinal study of 30,239 African-American and white adults aged  $\geq 45$  years found in the Reasons for Geographic and Racial Differences in Stroke (REGARDS) Study. The overall objective of the REGARDS study is to determine the causes for the excess stroke mortality in the Southeastern US and among African-Americans. Using geographically weighted regression methods, we were able to measure and map the local estimates of the racial differences in CVH between blacks and whites across the entire US. We additionally created maps of the varying spatial uncertainty in the local estimates of these differences to better gauge the significance of these racial differences. As a result, we found significant spatial patterning in the black-white differences in CVH – even beyond the well-known Stroke Belt and Stroke Buckle (located in the Southeastern region of the country). All of the estimated differences indicated blacks consistently having diminishing CVH compared to whites – where this difference was largely noted in pockets of the Stroke Belt and Stroke Buckle, in addition to moderate/large disparities noted in the Great Lakes region, portions of the Northeast and along the West coast.

Conclusions:

Efforts to improve CVH and ultimately reduce disparities between blacks and whites require culturally competent methods, with a strong focus on geography-based interventions and policies.

**AUTHORS/INSTITUTIONS:** L.P. Tabb, A. Ortiz, L. McClure, Department of Epidemiology and Biostatistics, Drexel University, Philadelphia, Pennsylvania, UNITED STATES|S. Judd, The University of Alabama, Birmingham, Alabama, UNITED STATES|M. Cushman, The University of Vermont Medical Center, Burlington, Vermont, UNITED STATES|

**CONTROL ID:** 3358018

**TITLE:** Single-Step Mycobacterium Preparation System for Clinical Diagnosis Using Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry

**ABSTRACT BODY:**

**Abstract Body:** Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI TOF MS) has been used as a high-throughput, cost-efficient identifier of microorganisms. In clinical diagnosis, the mass spectrometry (MS) profiles of pathogenic mycobacteria have been analyzed by modified pattern matching algorithms. To acquire proteomic MS profiles from mycobacteria, complicated mechanical and chemical pretreatments should be preceded from clinical isolates. Typically, these pretreatments require well-trained technicians for complex, labor-intensive steps due to the characteristics of mycobacteria which have rigid and waxy cell walls. Unfortunately, the data acquired by the conventional "Beads-beating" method would yield insufficient sensitivity for accurate identifications. Here, we suggest a more convenient pretreatment system, "Mycoprep", which helps to conduct simplified pretreatments in a single device. It can reduce processing time from an hour to 5 minutes. A micro/nano-technology is applied to integrate multiple, complex experimental protocols into a simplified system, called "Mycoprep". In "Mycoprep", a sharp and dense micro-structure induces the cell fragmentation and an integrated microfluidic channel network allows the reagent manipulation for chemical reactions. To validate the feasibility of the "Mycoprep" in mycobacterium identification, we used Non tuberculous mycobacteria (NTM) causing severe respiratory disease, with the MS profile database and identification algorithm of "MicroIDSys". The spectrum quality of MS profiles was evaluated by signal processing and a new revised analytical algorithm for mycobacterium spectra. Then, a novel "weight" assignment to every (m/z) in the acquired MS spectra was introduced to get a high identification performance. In summary, we suggested "Mycoprep" with simplified processing steps and less labor-intensive microorganism pretreatments in MALDI TOF MS analysis along with a weighted (m/z) concept, which explains the role of each (m/z) to the each NTM species. MS profile analysis with "weighted" classification algorithm and the new pretreatment protocol "Mycoprep" were the key for the success of mycobacterium identification. Our findings can be extended to the accurate and rapid clinical diagnosis of infectious disease.

**AUTHORS/INSTITUTIONS:** I. Lee, Y. In, J. Lee, D. Kim, C. Ko, E. Jo, Nosquest Corporation, Seongnamsi, KOREA (THE REPUBLIC OF)|E. Jo, Asta Incorporation, Suwonsi, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3358352

**TITLE:** Resampling-Based Analysis of Multivariate Repeated Measures Designs with the R Package MANOVA.RM

**ABSTRACT BODY:**

**Abstract Body:** Nonparametric statistical inference methods for a modern and robust analysis of longitudinal and multivariate data in factorial experiments are essential for research. While existing approaches that rely on specific distributional assumptions of the data (multivariate normality and/or equal covariance matrices) are implemented in statistical software packages, there is a need for user-friendly software that can be used for the analysis of data that do not fulfill the aforementioned assumptions and provide accurate p value and confidence interval estimates. Therefore, newly developed nonparametric statistical methods based on bootstrap- and permutation-approaches, which neither assume multivariate normality nor specific covariance matrices, have been implemented in the freely available R package MANOVA.RM.

The package provides routines for the analysis of multivariate data (MANOVA) and repeated measures designs and can even handle the combination of both: multivariate repeated measures data. The methods are illustrated using a data example from psychology.

**AUTHORS/INSTITUTIONS:** S. Friedrich, Department of Medical Statistics, University Medical Centre Göttingen, Göttingen, GERMANY|M. Pauly, Faculty of Statistics, Technische Universität Dortmund, Dortmund, GERMANY|F. Konietzschke, Charité Universitätsmedizin Berlin, Berlin, GERMANY|

**CONTROL ID:** 3358398

**TITLE:** Discrete survival models with flexible link functions for age at first marriage among women in Swaziland

**ABSTRACT BODY:**

**Abstract Body:** This study explores the use of flexible link functions in discrete survival models through a simulation study and an application to the Swaziland Demographic and Health Survey (SDHS) data. The objective of the research study is to perform simulation exercises in order to compare the effectiveness of different families of link functions and to construct a discrete multilevel survival model for age at first marriage among women in Swaziland using a flexible link function. The Pareto hazard model, Pregibon and Gosset families of link functions were considered in models with and without unobserved heterogeneity. The Pareto model where the family parameter is estimated from the data was found to outperform the other models, followed by the Pregibon and the Gosset family of link functions. The results from both simulation study and real data analysis of the SDHS data illustrated that, misspecification of the link function causes bias on the estimation of results. This demonstrates the importance of choosing the right link. The findings of this study reveal that women who are highly educated, stay in the Manzini and Shiselweni region, those who reside in urban areas were more likely to marry later compared to their counterparts in Swaziland. The results also reveal that the proportion of early first marriages is declining since the difference among birth cohorts is found to be very high, with women of younger cohorts getting married later compared to older women.

**AUTHORS/INSTITUTIONS:** T. Nevhungoni, Statistics, University of Venda, Thohoyandou, Limpopo, SOUTH AFRICA|

**CONTROL ID:** 3358477

**TITLE:** Statistical inference methods for two cumulative incidences in the presence of competing risks

**ABSTRACT BODY:**

**Abstract Body:** Competing risks data arise frequently in clinical trials, and a common problem encountered is the overall homogeneity between two groups. In competing risks analysis, when the proportional subdistribution hazard assumption is violated or when two cumulative incidence function (CIF) curves cross, the most common currently used testing methods, e.g., the Gray test and the Pepe and Mori test, may have a significant loss of statistical testing power. In this article, we propose a testing method based on the absolute difference in the area under the CIF curves. This method captures the difference over the whole time interval for which survival information is available for both groups and is not based on any special assumptions regarding the underlying distributions. This method was also extended to test short-term or long-term effects. We also consider a combined test and a two-stage procedure based on this new method that considers all possible alternatives, and a bootstrap resampling procedure is suggested in practice to approximate its limiting distribution. An extensive series of Monte Carlo simulations is conducted to investigate the power and the type I error rate of the methods. And from our simulations, our proposed ABC, Comb and TS tests have a relatively high power in most situations. Besides, the methods are illustrated using three different datasets, namely, data from a pediatric cancer trial, a malignant melanoma trial and an acute lymphoblastic leukemia trial, that have different situations of CIFs.

**AUTHORS/INSTITUTIONS:** J. Lyu, J. Chen, Z. Chen, Department of Biostatistics, Southern Medical University, Guangzhou, CHINA|Y. Hou, Department of Statistics, Jinan University, Guangzhou, CHINA|

**CONTROL ID:** 3358944

**TITLE:** A Bayesian approach to sparse reduced-rank regression modeling with binary outcomes

**ABSTRACT BODY:**

**Abstract Body:** In multivariate regression analysis, sparse reduced-rank regression (SRRR) has emerged as a powerful tool for handling high-dimensional multivariate data. While there has been extensive research on SRRR for continuous outcomes, SRRR methods for binary outcomes are still limited. To fill this research gap, we develop a fully Bayesian method for sparse reduced-rank regression modeling with binary outcomes. A major difficulty that occurs in our fully Bayesian framework is that the dimension of parameter space varies with the selected variables and the reduced-rank. To address the varying-dimensional problem, we propose a new approximate Bayesian inference procedure based on low-rank approximation techniques. A key feature of our fully Bayesian method is that the model uncertainty associated with rank selection and variable selection is automatically integrated. The proposed method is examined via a simulation study and real data analysis.

**AUTHORS/INSTITUTIONS:** G. Goh, D. Yang, H. Wang, Department of Statistics, Kansas State University, Manhattan, Kansas, UNITED STATES]

**CONTROL ID:** 3358990

**TITLE:** A covariate-adjusted classification model for multiple longitudinal biomarkers

**ABSTRACT BODY:**

**Abstract Body:** The classification methods based on a linear combination of multiple biomarkers have been widely used to improve the accuracy in disease screening and diagnosis. However, it is seldom to include covariates such as gender and age at diagnosis into these classification procedures. It is known that biomarkers or patient outcomes are often associated with some covariates in practice; therefore, the inclusion of covariates may further improve the power of prediction as well as the classification accuracy. In this study, we focus on the classification methods for multiple longitudinal biomarkers adjusting for covariates. With the use of a natural cubic spline basis, each longitudinal biomarker can be characterized by spline coefficients with a significant data dimension reduction. Technically, the maximum reduction can be achieved by controlling the number of knots or degrees of freedom in the spline approach, and then the spline coefficients can be obtained by the ordinary least squares method. We propose a non-parametric two-stage method to combine all spline coefficients obtained from every longitudinal biomarker and further adjust for covariates. Specifically, the optimal linear combination of those spline coefficients can be acquired using a stepwise method without any distributional assumption by maximizing the corresponding AUC. Afterward, covariates are included for additional improvement in classification. The asymptotic properties can be shown easily with the maximum rank correlation estimators. For illustration, the proposed method is applied to the longitudinal data of Alzheimer's disease and the primary biliary cirrhosis data. We also conduct an extensive simulation study to assess the finite-sample performance of the proposed method for multiple longitudinal biomarkers.

**AUTHORS/INSTITUTIONS:** S. Yu, W. Hsu, Statistics, Kansas State University, Manhattan, Kansas, UNITED STATES|P. Li, Statistics, Tamkang University, New Taipei City, TAIWAN|

**CONTROL ID:** 3359874

**TITLE:** Estimation of average treatment effects among multiple treatment groups by using an ensemble approach

**ABSTRACT BODY:**

**Abstract Body:** In observational studies, generalized propensity score (GPS)–based statistical methods, such as inverse probability weighting (IPW) and doubly robust (DR) method, have been proposed to estimate the average treatment effect (ATE) among multiple treatment groups. In this article, we investigate the GPS-based statistical methods to estimate treatment effects from two aspects. The first aspect of our investigation is to obtain an optimal GPS estimation method among four competing GPS estimation methods by using a rank aggregation approach. We further examine whether the optimal GPS-based IPW and DR methods would improve the performance for estimating ATE. It is well known that the DR method is consistent if either the GPS or the outcome models are correctly specified. The second aspect of our investigation is to examine whether the DR method could be improved if we ensemble outcome models. To that end, bootstrap method and rank aggregation method are used to obtain the ensemble optimal outcome model from several competing outcome models, and the resulting outcome model is incorporated into the DR method, resulting in an ensemble DR (enDR) method. Extensive simulation results indicate that the enDR method provides the best performance in estimating the ATE regardless of the method used for estimating GPS. We illustrate our methods using the MarketScan healthcare insurance claims database to examine the treatment effects among three different bones and substitutes used for spinal fusion surgeries. We draw conclusions based on the estimates from the enDR method coupled with the optimal GPS estimation method.

**AUTHORS/INSTITUTIONS:** X. Yan, K. Kulasekera, M. Kong, Bioinformatics and Biostatistics, University of Louisville, Louisville, Kentucky, UNITED STATES|S. Datta, Biostatistics, University of Florida, Gainesville, Florida, UNITED STATES|

**CONTROL ID:** 3360170

**TITLE:** Prevalence and risk factors for subjective psychosomatic symptom in Japanese adolescents: baseline analysis from the “SPRAT” cluster randomized controlled trial

**ABSTRACT BODY:**

**Abstract Body:** **BACKGROUND:** Recently, we developed a school-based parent-participant lifestyle education program for reducing subjective psychosomatic symptom (SPS) scores of adolescents (SPRAT), and conducted a cluster randomised clinical trial (cRCT) with two intervention arms (SPRAT vs. usual school education) [1]. To examine the prevalence of SPS and the factors associated with SPS using the baseline assessments is the objective of this presentation.

**METHODS AND ANALYSES:** This is a 6-month cRCT. Study participants were adolescents (aged 12–14 years) and their parents/guardians in middle high school, Japan. The primary endpoint was the change from baseline SPS scores to those obtained after 6 months. Between-group differences was analysed following the intention-to-treat principle. The sample size required was determined based on the information needed to detect a difference in the primary outcome with a significance level of 5% and power of 80% (40 students per cluster, effect size= 0.3, ICC= 0.02). In total, participation by 28 schools (students: n=1,120) was needed. Summary statistics were calculated for SPS score and factors related to lifestyle (11 questions) at baseline. Several mixed effects models were used for the analyses.

**RESULTS:** Overall, 1,959 adolescents (935 boys and 1,024 girls) among 2,150 (30 schools) responded to the baseline questionnaire and written informed consent. The mean SPS scores were 21.8 (SD=7.7) for boys and 23.4 (SD=7.3) for girls. Having lifestyles such as “more than 6hours sleep”, “fast asleep at 12 AM (midnight)”, “not consumed snacks after 10 PM” were protective factors for SPS for both boys and girls. The details will be shown in the presentation.

**CONCLUSION:** Higher SPS was frequent among adolescents having the lifestyle related to sleep and snacks in this population. The SPRAT program is the first study involving parents for reducing SPS in middle school students in Japan. To show the effectiveness of the SPRAT on reducing SPS score is warranted.

Trial registration number: UMIN 000026715.

**REFERECE**

[1] Watanabe J, Watanabe M, Yamaoka K, Adachi M, Nemoto A, Tango T. *BMJ Open*. 2018; 16;8(2):e018938. doi: 10.1136/bmjopen-2017-018938.

**AUTHORS/INSTITUTIONS:** J. Watanabe, Department of Nutrition, Minami Kyushu University, Miyazaki, JAPAN|M. Watanabe, Showa Women's University, Tokyo, JAPAN|K. Yamaoka, Prefectural University of Kumamoto, Kumamoto, JAPAN|K. Yamaoka, M. Adachi, A. Nemoto, T. Tango, Graduate School of Public Health, Teikyo University, Tokyo, JAPAN|T. Tango, Center for Medical Statistics, Tokyo, JAPAN|

**CONTROL ID:** 3360222

**TITLE:** BAYESIAN SURVIVAL ANALYSIS OF CERVICAL CANCER USING WEIBULL REGRESSION MODEL

**ABSTRACT BODY:**

**Abstract Body:** We investigate Bayesian survival analysis of cervical cancer in addition to a simulation study by Approximating posterior summary using Laplace Approximation and Laplace Demon function with Posterior mode, Posterior mean, posterior standard deviation and quartiles by fitting Weibull. From the MCMC algorithm with 100,000 iterations, discarding 10,000 as burn-in and from censoring levels of 10%, 20%, 50% and 80%, where we ordered all 440 cervical cancer dataset from smallest to largest, found the relevant percentile for the censoring time and set the top 10%, 20%, 50% and 80%, respectively, as censored. Then, we fit new dataset to estimate the model and ran posterior predictive checks to explore the uncertainty about the predictions. The result revealed that censoring has an effect on the performance of the Weibull models, that is, as the proportion of censoring increases, poorer parameter estimates were obtained in terms of both bias and precision, depending on the “closeness” of the components. More so, the study indicated that when the amount of censoring is small, very little bias is likely to result and confirmed that an acceptable model of survival data can still be obtained with light censoring up to 20%.

**AUTHORS/INSTITUTIONS:** S.A. FOLORUNSO, DEPARTMENT OF STATISTICS, UNIVERSITY OF IBADAN, Ibadan, Oy, NIGERIA|

**CONTROL ID:** 3360313

**TITLE:** Frailty models for recurrent event data

**ABSTRACT BODY:**

**Abstract Body:** Time-to-event data analysis has a long tradition in statistical applications. Frailty models are often used in multivariate survival analysis to deal with clustered event times. Whereas the shared frailty model is widely applied, the correlated frailty model has gained attention because it elevates the restriction of unobserved factors to act similar within clusters. Estimating frailty models is not straightforward due to various challenges. Recurrent event data appear in many fields of science and are an important field for application of frailty models. Often such recurrent events are followed by repair action in technical applications or medical intervention in life science. A model to deal with recurrent event times for incomplete repair of technical systems is the trend-renewal process. It is composed of a trend and a renewal component. In the present paper, we use a Weibull process for both of these components. The model is extended to include a Cox type covariate term to account for observed heterogeneity. A further extension includes a frailty term to account for unobserved heterogeneity. We fit the extended version of the trend-renewal process to data of hospital readmission times of colon cancer patients for illustration. Further real data examples of correlated event time models will be presented and their advantages and limitations are discussed in detail.

Pietzner D., Wienke A. (2013) The trend-renewal process: a useful model for medical recurrence data. *Statistics in Medicine* 32, 142 – 152

Wienke A. (2010) *Frailty Models in Survival Analysis*. Chapman and Hall/CRC, Boca Raton

**AUTHORS/INSTITUTIONS:** A. Wienke, Institute of Medical Epidemiology, Biostatistics, and Informatics, Martin-Luther-University Halle, Halle, GERMANY]

**CONTROL ID:** 3360315

**TITLE:** ON MULTIVARIATE LOG CONCAVITY OF ELLIPTICAL DENSITY WITH APPLICATION TO POULTRY FEEDS DATA

**ABSTRACT BODY:**

**Abstract Body:** Many recent publications have witness lots of resurgence in the theory and applications of log concavity of probability density functions. This method has some important characteristics which makes it appealing in modelling such as the existence of the maximum likelihood and the class of log-concave contains most of the commonly used parametric class of distributions. But all results obtained so far are based on the assumption that you have a large sample size, which in many cases is not achievable for some intricacies such as the cost, volatility of some areas, time and other constraints like epidemic outbreak. The present paper addresses this issue by extending the work of Chang and Walther (2007) using a class of multivariate elliptically contoured density as an underlining distribution for log concavity, this density function has shape parameter which regulates its tails for “lightness” or “heaviness”. The covariance matrix retains its interpretation so that it can easily be structured for serial dependence and several levels of variance components. This distribution performs creditably well whether the sample size is small or large. Some simulation study was carried out to examine its goodness-of-fit and application in the discrimination and classification of poultry feeds data are also considered.

**AUTHORS/INSTITUTIONS:** A.A. OLOSUNDE, MATHEMATICS, OBAFEMI AWOLowo UNIVERSITY, Ile-Ife, Osun State, NIGERIA|

**CONTROL ID:** 3360680

**TITLE:** Estimation of Nature History of Cancer and Screening-Related Overdiagnosis

**ABSTRACT BODY:**

**Abstract Body:** Cancer screening can detect cancer that would not have been detected in a patient's lifetime without screening. Standard methods for analyzing screening data do not explicitly account for the possibility that tumors may remain latent indefinitely. We reviewed some statistical methods to estimate nature history of cancer and to extend these methods by considering cancers as a mixture of those that progress to symptoms (progressive) and those that remain latent (indolent). We derived the theoretical properties and conducted extensive simulation studies.

Simulations demonstrate satisfactory performance of the estimation approach to identify model parameters subject to precise specifications of input parameters and adequacy of interval case sample sizes. In application to four breast cancer screening trials, the estimated indolent fraction varies between 2% and 35% when assuming test sensitivity at 80% and varying specifications for the earliest time before the start of the trial that participants could plausibly have developed cancer. This talk will cover a few joint papers with colleagues.

**AUTHORS/INSTITUTIONS:** Y. Shen, Biostatistics, UT MD Anderson Cancer Center, Houston, Texas, UNITED STATES|

**CONTROL ID:** 3360729

**TITLE:** MATHEMATICAL MODELLING OF THE INFLAMMATORY PHASE OF SKIN WOUND HEALING IN RATS

**ABSTRACT BODY:**

**Abstract Body:** The skin wound healing is a complex process divided into three overlapping and interdependent phases (inflammatory, proliferative and remodelling). The inflammatory response must occur rapidly to avoid chronic inflammation and it depends on biochemical, molecular and cellular events. The effective crosstalk between leukocytes and cytokines (proinflammatory and anti-inflammatory) lead to correct healing of the lesions. We considered a system of ordinary differential equations to model the inflammatory phase of skin wound healing process under oleoresin and hydroalcoholic extract from *Copaifera langsdorffii* treatments. The model can exhibit two stable steady states corresponding to healthy or unhealthy skin, nevertheless this study has been concentrated in a parameter search to healthy state in order to verify the treatment efficiency comparing the results of the oleoresin against hydroalcoholic extract. Thus, we have analysed the roles among the main leukocytes (neutrophils and macrophages), present in the inflammatory phase, and the inflammatory cytokines: interleukin 6 (IL-6) and interleukin 10 (IL-10). The model solution reproduced the dynamics of the neutrophils and macrophages during inflammatory phase, however there was a lack between numeric and biological results suggesting the necessity to improve the model. One possible strategy to enhance this model is to consider the interaction between the pro-inflammatory cytokine and macrophages in the mathematical model.

**AUTHORS/INSTITUTIONS:** P.F. Mancera, BBVPZ, UNESP, Botucatu, São Paulo, BRAZIL|C. Pellizzon, Morfology, UNESP, Botucatu, BRAZIL|L. Gushiken, Biotechnology Posgraduate Program, UNESP, Botucatu, BRAZIL|M. Oliveira, Biometric Posgraduate Program, UNESP, Botucatu, BRAZIL|

**CONTROL ID:** 3361479

**TITLE:** Group-based Dual Trajectory Modeling Approach for Association between Depression and Anxiety

**ABSTRACT BODY:**

**Abstract Body:** Background: Depression and anxiety are two common mental health problems. The aimed of this study: (i) to develop group-based dual trajectory of depression and anxiety; (ii) to identify the interrelationship with other adjust predictors between the depression and anxiety trajectory groups.

Methods: The study involved 3983 elderly who were age 65 or older between 2008 and 2015 from the Korea Health Panel study. Group-based dual trajectory modeling (GBDTM) was applied to find the heterogeneity and pathway of depression and anxiety. Logistic regression were used to identify the association between depression and anxiety.

Results: Four trajectory groups were identified in depression and anxiety outcomes from the elderly sample.

Depression has: "low-flat (87.0%)", "low-to-middle (8.8%)", "low-to-high (1.3%)" and "high-stable (2.8%)" trajectory groups. Anxiety has: "low-flat (92.5%)", "low-to-middle (4.7%)", "high-to-low (2.2%)" and "high-curve (0.6%)" trajectory groups. Based on the "low-flat" depression group, "low-to-high" depression group was more often involved in "low-to-middle" anxiety group. Individuals in the "low-to-middle" depression group had high chance in "low-to-middle" anxiety group and "high-curve" anxiety group. The members in high-stable depression group were more likely to belong to "high-to-low" anxiety group and "high-curve" anxiety group. Simulation study is performed based on two correlated outcomes.

Conclusion: This study can be used to monitor the older adults who has comorbid depression and anxiety and also to assist health policy decision-makers for planning intervention programs.

**AUTHORS/INSTITUTIONS:** Y. Cheng, School of Public Health, University of Saskatchewan, Regina, Saskatchewan, CANADA|

**CONTROL ID:** 3362050

**TITLE:** Comparative Analysis of CNN Approaches for Fingerprint Anti-Spoofing

**ABSTRACT BODY:**

**Abstract Body:**

In order to increase the security of fingerprint recognition, anti-spoofing methods in fingerprint recognition systems are crucial. This paper contributes to finding a reliable and practical anti-spoofing method using convolutional neural networks (CNN). Among various models of CNN, this paper compares six models: LeNet, AlexNet, GoogLeNet, VggNet, ResNet, and DenseNet, in order to find the most proper models for fingerprint anti-spoofing. The model with the highest average accuracy is reinforced by the change in various parameters. In addition to the highest generalization performance, this paper also contains the models with high accuracy associated with parameters and mean average error rates to find the model that consumes the least memory and computation time for training and testing. Among the CNN models, AlexNet has the overall highest accuracy in our anti-spoofing fingerprint database. AlexNet with 6 layers, no data augmentation, and uses average pooling layer shows the average highest accuracy, with an accuracy of 96.19%, memory usage of 3.13MB, and mean average error rate of 0.06. On the other hand, data augmentation lowers the accuracy and increases the mean average error rate, so data augmentation may influence negative impacts on many of CNN models. More wide usage of data augmentation strategy needs to be investigated. Although Alexnet shows the highest performance with moderate use of memory, more research is needed to increase the accuracy of the CNN models for anti-spoofing purposes, in terms of loss functions, optimization techniques, and generalization power.

**AUTHORS/INSTITUTIONS:** Y. heo, Turlock Christian High School, Yongin, Gyeonggi-do, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3362186

**TITLE:** ANALYSIS OF STUDENT LECTURER INTERACTION IN ONLINE PHT 112 COURSE AT MASENO UNIVERSITY.

**ABSTRACT BODY:**

**Abstract Body:**

Introduction

**Area of study**The study was based in Maseno university main campus. Maseno University comprises a total of approximately 15,000 students of whom we have both the Self sponsored students and Government sponsored students. Elearning in the University is managed and supervised using a Learning Management System (LMS Moodle) module in the E campus. The E campus offices are situated at Maseno University Varsity Plaza in Kisumu city, Kenya. **Target Population**The population comprised all the Maseno university students in the main campus. It is comprised of first years, second years, third years and fourth years for it is compulsory in Maseno university to undertake PHT 112 online courses since the registration of the course is determined by the student furthermore it does not exceed the fourth year of study.

Sample design

Simple random sampling was used to draw 2 groups from the total 10 groups in LMS module. 10 topics for both groups were compiled together. The topics were further divided into six event activities that took place in the discussion forum which included: discussions creation(by students), posts creation(by students), course module updating(by the lecturers), discussion views(by students) and contents in each discussion The lecturers involved were 5 who facilitated the two groups. The event activities were further ranged in months and the total in each month was calculated.

Data collection method

Our project involved secondary data which was sourced from the LMS module of Maseno E campus. The data was collected by means of a flash disc drive which we transferred in our PC's for further analysis.

Data analysis

Data collected was analyzed using Descriptive statistics (bar graphs) and inferential statistics (Non parametric tests, Q Q plots and histograms). The above was achieved using SPSS and Excel statistical softwares.

**AUTHORS/INSTITUTIONS:** M.D. Wanjiru Muthoni, Pure and applied Mathematics, Maseno University, Nairobi, KENYA|

**CONTROL ID:** 3362676

**TITLE:** Pairwise joint modelling of clustered and high-dimensional pneumonia care outcomes with covariate missingness

**ABSTRACT BODY:**

**Abstract Body:** Background: Multiple outcomes reflecting different aspects of routine care are a common phenomenon in health care research. A common approach of handling such outcomes is multiple univariate analyses. Alternatively, the outcomes are combined into a single composite outcome. Whilst these approaches are straight forward, they do not provide an avenue for answering research questions such as associations amongst the outcomes, and joint effects of covariates. In this work, we sought to study associations amongst paediatric pneumonia care outcomes, while circumventing the computational challenge posed by their clustered and high-dimensional nature, and accounting for missing covariates.

Methods: Data were from a cluster randomized trial investigating the effects of audit and feedback on quality of clinicians' prescribed routine paediatric care in 12 Kenyan hospitals. The outcomes of interest were 9 binary pneumonia care indicators spanning three domains of care namely assessment, diagnosis and treatment of pneumonia patients. There were varying degrees of missingness in the covariates of interest, and these were imputed using the latent normal joint modelling multiple imputation method. To estimate the joint model for the 9 binary outcomes, we used the pairwise joint modelling strategy to fit all 36 bivariate random intercepts models and combine inferences using pseudo-likelihood theory.

Results: Preliminary results indicated varying strength and direction of association amongst pneumonia outcomes within and across the 3 domains of pneumonia care. While some outcomes in the assessment domain exhibited positive association, others showed negative association, with correlation of the respective random intercepts ranging from -0.02 to -0.71. In the treatment domain, two outcomes namely prescription of oral amoxicillin and correct dosage were positively associated, with a corresponding correlation of 0.795. Across domains of care, pneumonia diagnosis was strongly correlated with both outcomes in the treatment domain, but this was not the case with outcomes in the assessment domain.

**AUTHORS/INSTITUTIONS:** S. Gachau, M. English, Health Services Unit, KEMRI-Wellcome Trust Reserach Programme and University of Nairobi, Nairobi, Nairobi, KENYA|S. Gachau, N. Owuor, School of Mathematics, University of Nairobi, Nairobi, KENYA|E. Njagi, Department of Non-Communicable Disease Epidemiology,, London School of Hygiene and Tropical Medicine, London, UNITED KINGDOM|G. Molenberghs, Interuniversity Institute for Biostatistics and statistical Bioinformatics,, Universiteit Hasselt,, Diepenbeek, BELGIUM|G. Molenberghs, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Katholieke Universiteit Leuven, Leuven, BELGIUM|M. English, Nuffield Department of Medicine, University of Oxford, United Kingdom, Oxford, UNITED KINGDOM|P. Ayieko, Mwanza Intervention Trials Unit, Mwanza, TANZANIA, UNITED REPUBLIC OF|P. Ayieko, Department of Infectious Disease Epidemiology,, London School of Hygiene and Tropical Medicine,, London, UNITED KINGDOM|

**CONTROL ID:** 3363160

**TITLE:** Addressing data privacy in dynamic treatment regimen estimation via virtual pooling

**ABSTRACT BODY:**

**Abstract Body:** Personalized medicine is a rapidly expanding area of health research wherein patient level information is used to inform treatment decisions. Dynamic treatment regimens (DTRs) provide a statistical framework to formalize the individualization of treatment decisions that characterize personalized management plans. Numerous methods have been proposed to estimate DTRs that optimize expected patient outcomes, many of which have desirable properties such as robustness to model misspecification. However, while individual data are essential in this context, there may be concerns about data confidentiality, particularly in multi-centre studies where data are typically shared externally. To address this issue, we adopted the privacy preserving approach of data pooling, and considered the application of a robust yet user-friendly estimation approach, dynamic weighted ordinary least squares, to these data. In simulations, we evaluated the performance of the method in estimating the parameters of the decision rule under different assumptions concerning (i) the distribution of the subject-specific covariates in each pool, (ii) the treatment variable type (binary vs dosage), (iii) the strength of the relationship between the treatment and the subject-specific covariate, (iv) the pool size when aggregating the data, and (v) the model misspecification. The results demonstrated that in the data pooling setting, the treatment model is generally not correctly specified and hence double robustness is lost. This can result in bias, which can be low, moderate or high, depending on the different combination of the five points described above. We illustrated this finding on an example of warfarin dosing using data from the International Warfarin Consortium data.

**AUTHORS/INSTITUTIONS:** C. Danieli, E. Moodie, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, CANADA]

**CONTROL ID:** 3363739

**TITLE:** Using linear and non linear functional additive effects for predicting binary variables: an Bayesian approach with an application to EEG data

**ABSTRACT BODY:**

**Abstract Body:** We propose a two level Bayesian hierarchical model. In the first level, we have a gaussian model and in the second level we have a binary response model. The response for the first level is a known gaussian outcome, however the binary response involved in the second level is not observed. Therefore, we need to predict its value before modeling its probability by means of a GLM model. We consider two different GLM models for the binary outcome: one considers scalar and linear functional effects, while the other considers scalar and non linear functional effects. We perform a Bayesian analysis for fitting models, hence the non observable binary variable come into the context of estimation as an unknown. By using a Gibbs sampling algorithm type we are able to predict the latent binary response by sampling from its full conditional distribution. We expand the linear and non linear functional effects by using cubic B-splines and associate a prior distribution to the coefficients in the expansion. We then sample from the posterior distribution of the expansion coefficients following Gelman, et al. [The annals of applied statistics, 2, 1360-1386 (2008)]. We work with a data set that resulted from a randomized placebo-controlled trial which has been studied by Jiang et al. [The annals of applied statistics, 11.3, 1513-1536 (2017)]. Based on these data, the goal is to identify a subgroup of subjects, named early responders, who experience symptoms improvement early on during antidepressant treatment by using the information of fourteen electroencephalograms (EEGs), which is considered to be an indication of a placebo instead of a true pharmacological response. We estimate the effects of the fourteen EEGs in linear and non linear functional approach in the context of GLM regression and were able to predict those subjects who may have experienced symptoms improvement early on during antidepressant treatment by identifying four EEGs whose effects are important to predict the probability of being an early responder.

**AUTHORS/INSTITUTIONS:** M.R. Motta, N.L. Garcia, Statistics, University of Campinas, Campinas, Sao Paulo, BRAZIL|E. Petkova, T. Tarpey, New York University, New York, New York, UNITED STATES|T. Ogden, Columbia University, New York, New York, UNITED STATES|

**CONTROL ID:** 3364081

**TITLE:** Two-phase Design and Analysis in Post-Genome-Wide Association Studies: Challenges and Opportunities

**ABSTRACT BODY:**

**Abstract Body:** Two-phase design and analysis is a cost-reduction technique that collects expensive molecular data (G) in an informative subset of a large genome-wide association study (GWAS) cohort for post-GWAS investigation. Recently, we developed methods to identify optimal two-phase designs for post-GWAS targeted sequencing analysis under a semi-parametric maximum likelihood (SPML) framework within the exponential family thereby improving cost-efficiency under budgetary limitations. Here we report evaluation of hypothesis testing for association with a continuous outcome. At phase 1, a GWAS on a quantitative trait, Y, determines a genomic region of interest with a corresponding top-GWAS SNP, Z, as an auxiliary covariate. In phase 2, G is measured on a subsample selected according to values of Y and Z, making G missing by design and reducing the total cost of molecular technologies. Lastly, inference on G via SPML efficiently utilizes available data from phases 1 and 2.

Given design quantities, i.e. an effect size for G, and a G-Z haplotype distribution, we propose two approaches: (1) a Laplace multiplier (LM) method based on an analytical expression for the variance-covariance matrix (VCM) subject to a budget constraint, and (2) a genetic algorithm (GA) that performs a direct search of the phase 2 discrete space of potential subsamples. We evaluate VCM optimality criteria useful in experimental design. Comprehensive simulation studies to assess the empirical properties of LM and GA suggest that the optimal designs can render higher power compared to heuristic designs while preserving type 1 error.

The main challenge of implementing these optimal designs, however, arises in specification of unknown design quantities. We evaluate a strategy to select a phase 2 subsample at the design stage in which investigators postulate a range of hypothesized design factors (genetic effects, minor allele frequencies, linkage disequilibrium patterns). Information across the postulated factors is combined under the proposed designs by a min-median approach. We illustrate this procedure in the Northern Finland Birth Cohort 1966 comprised of 5402 subjects from Finland's two northernmost provinces. We demonstrate that two-phase studies can drastically reduce the costs of gathering molecular data without substantial loss of power to detect genetic associations.

**AUTHORS/INSTITUTIONS:** O. Espin-Garcia, Biostatistics, Princess Margaret Cancer Centre, Toronto, Ontario, CANADA|O. Espin-Garcia, S.B. Bull, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, CANADA|R. Craiu, Statistical Sciences, University of Toronto, Toronto, Ontario, CANADA|S.B. Bull, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, CANADA|

**CONTROL ID:** 3364694

**TITLE:** A flexible method for fitting and forecasting mortality rates using Bernstein polynomials

**ABSTRACT BODY:**

**Abstract Body:** Analyses of mortality data have played a significant role on the fields such as epidemiology, medicine and public health. Most of the existing mortality models mainly focus on providing a good fitting, to the damage to a perfect forecasting result. In this investigation, we propose a flexible maximum likelihood approach for modeling and forecasting the mortality rates by using the two-dimensional Bernstein polynomials. We use Bernstein coefficients to construct a regression model with shape restriction and use the simulated annealing method to perform the maximum likelihood based inference. The proposed method is intuitive and use the characteristic of Bernstein polynomials to describe the pattern of mortality trends. The performance of the proposed method is examined by simulated examples and illustrated through the practical dataset. The simulated examples and the practical data analysis show that the proposed method has better performance than the existing mortality models in forecasting the short-term mortality trends.

**AUTHORS/INSTITUTIONS:** L. Chien, Center for Fundamental Science, Kaohsiung Medical University, Kaohsiung, TAIWAN|Y. Wu, L. Hong, L. Cheng, Chung Yuan Christian University, Taoyuan, TAIWAN|

**CONTROL ID:** 3364769

**TITLE:** Modelling dependence structures of extreme wind speed using bivariate distribution: a Bayesian approach

**ABSTRACT BODY:**

**Abstract Body:** When investigating extremes of weather variables, it is often not just a single station which determines the damage caused, but in turn extremes may be caused from the combined behaviour of several weather stations. In order to investigate the joint dependence of extreme wind speed, a bivariate generalised extreme value distribution (BGEVD) have been considered from the frequentist and Bayesian approaches to analyse the extremes of component wise monthly maximum wind speed at selected weather stations in South Africa. In the frequentist approach, the parameters of EVDs were estimated using maximum likelihood, whereas in the Bayesian approach the Markov Chain Monte Carlo technique with the Metropolis-Hastings algorithm was used. The results show that the BGEVD tted to component wise maxima of extreme weather variables provide apparent benets over the univariate method, which allows information to be pooled across stations and resulted improved precision of the estimate for the parameters as well as return levels of the distributions.

**AUTHORS/INSTITUTIONS:** T.A. Diriba, L. Debusho, Statistics, University of South Africa, Johannesburg, SOUTH AFRICA|

**CONTROL ID:** 3365811

**TITLE:** Selection of genotypes according to quality parameters in comparative multi-environmental trials of cultivars.

**ABSTRACT BODY:**

**Abstract Body:** The quality of a genotype (G) is defined by multiple variables specific to each cultivar. The selection of a wheat cultivar (*Triticum aestivum* L.) for its quality attributes is key to defining the destination of the harvest. However, these attributes vary in different degrees depending on the environmental (E) effect and the genotype interaction by environment (GxE), making such a selection difficult. In this work we present a method to differentiate commercial cultivars groups according to quality parameters. The methodology implies (i) estimation of factors with the maximum amount of information from correlated variables, (ii) identification of quality groups based on quality parameters and (iii) estimation of the genetic contribution to intragroup variability for each group of quality. GxE matrices of quality data from 235 wheat genotypes, grown in 19 environments, divided into six trigerous subregions of Argentina, were analyzed for ten years. Nine important quality attributes in commercialization were used to determine the quality subgroups. The proposed methodology represents an approach to the selection of commercial genotypes, and optimization in terms of the effect on the selection and sampling of genotypes in multi-environmental trials across subregions.

**AUTHORS/INSTITUTIONS:** E. Del Vecchio, M. Balzarini, biometric, Universidad Nacional de Cordoba, Cordoba, ARGENTINA|P. Abbate, Intituto Nacional Tecnologico de Agricultura, Buenos Aires, ARGENTINA|

**CONTROL ID:** 3365970

**TITLE:** Body Mass Index and Systolic Blood Pressure impact on survival of dialytic patients: a multivariate joint model approach

**ABSTRACT BODY:**

**Abstract Body:** Context: Studies on impact of routine clinical parameters on mortality of patients on peritoneal dialysis (PD), especially among elderly, are inconsistent. Although association between body mass index (BMI) and systolic blood pressure (SBP) is well known, their joint role on PD patient evolution is not clear. Objective: To model the joint impact of BMI and SBP trajectories on a cohort of incident elderly PD patients. Methods: This was a prospective multicenter cohort study (Dec /2004 - Oct /2007) comprising 674 patients. Socio-demographic and clinical data were evaluated on patients followed until death. The modelling framework for the repeated measurements is the multivariate linear model with random effects and/or correlated error structure. The model for the time-to-event outcome is a Cox proportional hazards model with log-Gaussian frailty. Stochastic dependence is captured by allowing the Gaussian random effects of the linear model to be correlated with the frailty term of the survival model. This correlation was assessed using the random intercepts of BMI and SBP, and/or their random slopes as well as its current value, allowing for different clinical interpretations of this impact in the patients survival. All sub-models were adjusted by age in years, given the potential impact of this variable on BMI, SBP and death. The package JoinerML [1] were used for estimation. Results: Malnourished patients presented higher percentage of death compared to overweight and obese (44.6%, 26% and 29%,  $p = 0.001$ ) and lower BPS (129 mmHg, 142mmHg and 142 mmHg). Isolated current values effects of increased BMI and BPS showed both as protective factors (1% and 2% risk reduction, respectively, for each unitary increase), but when both were jointly modeled BMI changed its effect to 5% risk reduction while BPS remained the same, showing great impacts on survival depending on the combination of trajectories. Conclusions: Clinical parameters association and its impact on survival is complex, and estimation of isolated effects may provide biased survival curves leading to unrealistic prognosis. A multivariate joint modeling can be seen as a valuable tool for clinical protocols on patient evolution.

**AUTHORS/INSTITUTIONS:** F. Colugnati, N. Fernandes, Federal University of Juiz de Fora, Juiz de Fora, BRAZIL|

**CONTROL ID:** 3365981

**TITLE:** Use of mixed non-linear models in the estimation of leaf litter decomposition rates

**ABSTRACT BODY:**

**Abstract Body:** Nutrient cycling is an ecosystem process associated with soil fertility that depends on the decomposition of the organic matter. Decomposition is a biological process that studies how, organic matter, returns to the soil as nutrients. The most frequently used methodology in the study of nutrient cycling is based on estimates the leaf litter decomposition rate ( $k$ ) over time. The analyzed variable is the difference between the initial content of the leaf litter contained in bags versus the amount of litter that remains in the bags after a given time, expressed as a percentage. The leaf-litter bags are designed to allow the entry of decomposing microorganisms.

The classic statistical modeling strategy to estimate  $k$  from the percentage of remaining material on leaf-litter bags, consists in adjusting an exponential model using the set of bags evaluated over time for each treatment and each repetition. These estimates are then used to perform an ANOVA to assess differences between treatments. The data set used in this study came from 4 sites in an altitudinal gradient in Costa Rica, in each of which, 24 sets of bags were located (4 sites by 6 repetitions). Each set contained 10 bags, which were processed, one at a time, in 10 sampling moments. The same experimental design was used to evaluate leaf-decomposition rate for 10 different tree species. Our proposal is based on adjusting a mixed non-linear model for each species including the random effect of the repetition as well as the fixed effect of site on the  $k$  parameter. The comparison between both strategies was made by comparing the standard errors of the estimated means  $k$  by site and species. The use of mixed non-linear models led to estimates with lower standard error.

**AUTHORS/INSTITUTIONS:** E.E. Corrales, F. Casanoves, CATIE - Centro Agronómico Tropical de Investigación y Enseñanza, Turrialba, Cartago, COSTA RICA|J. Di Rienzo, Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, Córdoba, ARGENTINA|

**CONTROL ID:** 3367177

**TITLE:** A unified genetic association test robust to latent population structure

**ABSTRACT BODY:**

**Abstract Body:** We present a new statistical test of association between a trait and genetic markers, which we theoretically and practically prove to be robust to arbitrarily complex population structure. The statistical test involves a set of parameters that can be directly estimated from large-scale genotyping data, such as those measured in genome-wide associations studies. We also derive a new set of methodologies, called a genotype-conditional association test, shown to provide accurate association tests in populations with complex structures, manifested in both the genetic and non-genetic contributions to the trait. We demonstrate the proposed method on a simulation study and on the real data. Our proposed framework provides a substantially different approach to the problem from existing methods.

**AUTHORS/INSTITUTIONS:** M. Song, Department of Statistics, Sookmyung Women's University, Seoul, KOREA (THE REPUBLIC OF)

**CONTROL ID:** 3367234

**TITLE:** Nonlinear and spatially adjusted modelling of fertility using Nigeria demographic and health surveys, 2003–2013

**ABSTRACT BODY:**

**Abstract Body:** Background: One of the key indicators of size and composition of population is fertility. Fertility rate determines the family size; the freedom to make choices on spacing; timing; and number of desired pregnancies, which can influence the health, education, income and living conditions of an individual or family.

**Methods:**

The objective of the study is to investigate associated factors that contributes to fertility preferences in Nigeria. Specifically, this study determines the spatio-temporal variations in fertility between 2003 and 2013 that can be attributed to changing characteristics of women aged between 15 to 49 years using Nigeria Demographic and Health Survey (NDHS) data. We used Geo-additive mixed models (GAMM) to estimate the nonlinear association between number of children ever born (TCEB) and age (years) and region-specific terms. The nonlinear effects of age was modelled via cubic spline with three knots placed at equidistance and reference was set at mean age. The region-specific terms were treated as correlated random effects.

**Results:**

There was a decrease in percentage of women who never born children from 32.9% in 2003 to 29.5% in 2013. It was equally reported that women's age at their first marriage were (76.3% in 2003, 72.5% in 2008, and 70.3% in 2013) being 19 years of age and younger. In the 2008 and 2013 survey years, there were significant association between TCEB and the knowledge about contraceptives ( $p < 0.001$ ); contraceptive use ( $p < 0.0001$ ); and ideal family size ( $p < 0.001$ ). In the 2008 NDHS, there were significant association between TCEB and the religion ( $p < 0.01$ ); and women's age at first sex ( $p < 0.001$ ) with an odds ratio (OR) of 1.071 and 0.902. Similarly, in 2013 NDHS, there were significant association between TCEB and the religion ( $p < 0.001$ ); woman's age at first marriage ( $p < 0.001$ ); and woman's age at first sex ( $p < 0.001$ ).

**Conclusion:**

Women of the following characteristics had a greater fertility with the total children ever born: women whose place of residence are in the northwest and northeast geopolitical regions, aged between 20 – 29 years and 30 – 39 years at their first marriage, had their first sex with their partner at the age between 16 – 19 years with the notion that the ideal family size is at least 5, having at least 30 years old as the household head, and were Muslims and Christians.

**AUTHORS/INSTITUTIONS:** L.O. Mashood, W.B. Yahya, Statistics, University of Ilorin, Ilorin, Kwara State, Nigeria., Ilorin, Kwara State, NIGERIA|O.A. Adegboye, Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam, Ho Chi Minh, VIET NAM|O.A. Adegboye, Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh, VIET NAM|M.A. Adegboye, American University of Nigeria, 640001, Yola, Adamawa, NIGERIA|F.M. Elfaki, Department of Mathematics, Statistics and Physics, Qatar University, Doha, 2713, QATAR|

**CONTROL ID:** 3367236

**TITLE:** Imputing missing values in multi-environment trials using the singular value decomposition: An empirical comparison

**ABSTRACT BODY:**

**Abstract Body:** Missing values for some genotype-environment combinations are commonly encountered in multi-environment trials. The recommended methodology for analyzing such unbalanced data combines the Expectation-Maximization (EM) algorithm with the additive main effects and multiplicative interaction (AMMI) model. Recently, however, four imputation algorithms based on the Singular Value Decomposition of a matrix (SVD) have been reported in the literature (Biplot imputation, EM+SVD, GabrielEigen imputation, and distribution free multiple imputation - DFMI). These algorithms all fill in the missing values, thereby removing the lack of balance in the original data and permitting simpler standard analyses to be performed. The aim of this paper is to compare these four algorithms with the gold standard EM-AMMI. To do this, we report the results of a simulation study based on three complete sets of real data (eucalyptus, sugar cane and beans) for various imputation percentages. The methodologies were compared using the normalised root mean squared error, the Procrustes similarity statistic and the Spearman correlation coefficient. The conclusion is that imputation using the EM algorithm plus SVD provides competitive results to those obtained with the gold standard. It is also an excellent alternative to imputation with an additive model, which in practice ignores the genotype-by- environment interaction and therefore may not be appropriate in some cases.

**AUTHORS/INSTITUTIONS:** M. García Peña, Matemáticas, Pontificia Universidad Javeriana, Bogotá, COLOMBIA|S. Arciniegas Alarcón, Matemáticas, Universidad de La Sabana, Bogotá, COLOMBIA|W.J. Krzanowski, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UNITED KINGDOM|C. dos Santos Dias, Ciências Exatas, Universidade de São Paulo, Piracicaba, BRAZIL|

**CONTROL ID:** 3367239

**TITLE:** An alternative methodology for imputing missing data in trials with genotype-by-environment interaction: some new aspects

**ABSTRACT BODY:**

**Abstract Body:** A common problem in multi-environment trials arises when some genotype-by-environment combinations are missing. In Arciniegas-Alarcón et al. (2010) we outlined a method of data imputation to estimate the missing values, the computational algorithm for which was a mixture of regression and lower-rank approximation of a matrix based on its singular value decomposition (SVD). In the present paper we provide two extensions to this methodology, by including weights chosen by cross-validation and allowing multiple as well as simple imputation. The three methods are assessed and compared in a simulation study, using a complete set of real data in which values are deleted randomly at different rates. The quality of the imputations is evaluated using three measures: the Procrustes statistic, the squared correlation between matrices and the normalised root mean squared error between these estimates and the true observed values. None of the methods makes any distributional or structural assumptions, and all of them can be used for any pattern or mechanism of the missing values.

**AUTHORS/INSTITUTIONS:** S. Arciniegas Alarcón, Matemáticas, Universidad de La Sabana, Bogotá, COLOMBIA|M. García Peña, Matemáticas, Pontificia Universidad Javeriana, Bogotá, COLOMBIA|W.J. Krzanowski, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UNITED KINGDOM|C. dos Santos Dias, Ciências Exatas, Universidade de São Paulo, Piracicaba, BRAZIL|

**CONTROL ID:** 3367241

**TITLE:** Iron Status and Iron Regulatory Hormones in Hemochromatosis: Hepcidin and Erythroferrone

**ABSTRACT BODY:**

**Abstract Body:** The iron overload disorder, hereditary hemochromatosis (HH) occurs in approximately 5 persons per 1,000 in northern European populations. Most HH patients are homozygous for the p.c292Y (rs1800652) mutation in the hemochromatosis gene (HFE), leading to deficiency of hepcidin, the iron-regulatory hormone that normally controls dietary iron absorption. The hormone erythroferrone (Erfe), produced by marrow erythroblasts, also reduces hepcidin production and is a candidate modulator of HH severity. We analyzed 336 serum samples from the NIH BioLincc biorepository from Caucasian p.C282Y homozygotes (HH patients, N=186) and control participants (N=150) in the NIH Hemochromatosis and Iron Overload Screening (HEIRS) Study. The relationships of hepcidin and Erfe concentrations with iron overload in this cohort were determined by linear regression analysis. In the entire cohort, our analysis revealed a strong positive relationship between serum hepcidin concentration and iron stores assessed as natural log of serum ferritin, Ln(SF) ( $p < 0.0001$ ). In contrast, there was a modest negative relationship between Erfe concentration and Ln(SF) ( $p = 0.0479$ ). Analyses of the relationships between hepcidin and iron stores in HH patients and controls revealed a constant difference such that, on average, the expected mean value of hepcidin in the patients was 22.5 ng/ml lower than in controls ( $p = 2.00 \times 10^{-14}$ ). The positive relationship between hepcidin and Ln(SF) in this cohort is consistent with the known increase of hepcidin levels with higher body iron stores. However, the fact that the expected mean value of hepcidin in HH patients was significantly lower on average than in controls for any given value of Ln(SF) indicates a relative hepcidin deficiency in HH regardless of iron stores. The modest inverse relationship between Erfe levels and Ln(SF) in the current cohort may reflect increased Erfe production in response to stimulation of red cell production by phlebotomy therapy to remove excess iron. Further research is needed to determine whether extremely low hepcidin levels or elevated Erfe levels may predict the risk of severe iron overload in younger HH patients who have not yet accumulated increased iron stores.

**AUTHORS/INSTITUTIONS:** C.E. McLaren, G.D. McLaren, Department of Medicine, University of California, Irvine, Irvine, California, UNITED STATES|W. Chen, University of California, Irvine, Orange, California, UNITED STATES|T. Ganz, E. Nemeth, University of California, Los Angeles, Los Angeles, California, UNITED STATES|

**CONTROL ID:** 3367277

**TITLE:** Combination of different methodologies for climate time series forecasting

**ABSTRACT BODY:**

**Abstract Body:** Time series analysis is of fundamental importance for many areas of knowledge. There are currently a wide variety of techniques for forecasting time series. Traditionally, the Autoregressive Integrated Moving Average (ARIMA) model has been one of the most widely used linear models for time series prediction, since it can model and predict future values based on previous observations with some precision. However, a number of more sophisticated techniques and models have emerged. The combination of forecasts allows you to increase the accuracy of the final forecast in time series. This work aimed to explore and predict the future values of a time series by combining different models, using: ARIMA and MLP / RNA (Multilayer Perceptron Artificial Neural Network), Wavelet Combination with Multiple Stages, Hybrid model ARIMA-SVM (Support Vector Machine) and lastly the LSTM (Long Short Term Memory) model which is a type of recurrent neural network, suitable for classification, processing and prediction in time series with time intervals. Unknown the different methodologies adopted in the study were used to forecast climatic time series. The accuracy measures used to compare the models were RMSE and MAPE. According to the results obtained it was found that the application of the combined forecasting methods obtained better performance compared to the individual models.

Keywords: Hybrid models, Machine Learning, Neural networks, Deep learning, Wavelets.

**AUTHORS/INSTITUTIONS:** P.S. Guimarães, Departamento de Estatística, Universidade Federal de Lavras, Lavras, Minas Gerais, BRAZIL|

**CONTROL ID:** 3367309

**TITLE:** Estimation of Association and Minimum Detectable Association (MDA) between Two Types of Cancer Deaths

**ABSTRACT BODY:**

**Abstract Body:** Eiji Nakashima

Research Institute for Radiation Epidemiology and Biostatistics (RIREB)

Aki-gun Fuchu-cho Osu 1-6-28-505, Hiroshima 735-0021, Japan.

(Formerly a member in Statistics of Radiation Effects Research Foundation, Hiroshima Japan)

**ABSTRACT:** Radiation biology experimental evidences can be epidemiologically examined in atomic-bomb survivor cohort, if it is mathematically formulated adapting to the grouped data. This paper gives methods for MDA (Minimum Detectable Association) between two cancer mortalities if no joint observations of two cancer death and for the association if we observe the number of joint two cancer deaths. These can show the existence of "bystander effect" or genomic instability in radiation biology. Radiation induced genomic instability is a common cause of carcinogenesis resulting the association of two cancers death. We also mention the proneness of the multiple cancer in an exposed subjects using the correlation constructed by association and marginal hazards.

When no dose error, reasoning are based on following findings, A): Inflammatory diseases is one of the causes of cancers. B): Radiation induces subclinical persistent (long-term) inflammation (Neriishi, Nakashima & DeLongchamp. 2001, Int J Rad Biol) after adjusting the confounders. C): Radiation has direct carcinogenic effect (Ozasa et al. 2012, Rad Res). D): Radiation induced genomic instability is experimentally examined (Wright 1998 & his other papers, and Ullrich 1998, Int J Rad Biol). E): Radiation induced genomic instability might be dominant (large attributable fraction) in carcinogenesis (Ohtaki & Niwa 2001, Rad Res). Issues A) & B) indicate radiation related (non-decreasing) association between two cancers death. The unsolved problems are: Biological mechanisms are difference between childhood (acute) and adult cancers, and between hematopoietic and other cancers. On this problem, explanation by radiation induced genomic instability can explain major/some fraction of of radiation induced carcinogenesis. Note that dose error can change (linear-)dose response (Eiji Nakashima, Communication in Statistics, 2019, Vol.48(22):5517-5529) and the association. This poster includes further suggestions of my paper in the proceedings of the Japanese Society of Computational Statistics (JSCS), held in Tokyo (Aoyamagakuin University), Nov30-Dec 1, 2019.

**AUTHORS/INSTITUTIONS:** E. Nakashima, Research Institute for Radiation Epidemiology and Biostatistics, Hiroshima prefecture, JAPAN|

**CONTROL ID:** 3367323

**TITLE:** Farm-data analysis: Increased age at first-mating interacting with herd size or herd productivity decreases longevity and lifetime reproductive efficiency of sows in breeding herds

**ABSTRACT BODY:**

**Abstract Body:** The objectives of the present study were to characterize sow life and herd-life performance and examine two-way interactions between age at first-mating (AFM) and either herd size or herd productivity groups for the performance of sows. Data contained 146,140 sows in 143 Spanish herds. Sow life days is defined as the number of days from birth to removal, whereas the herd-life days is from AFM date to removal date. Herds were categorized into two herd size groups and two productivity groups based on the respective 75th percentiles of farm means of herd size and the number of piglets weaned per sows per year: large (1,017 sows or more) or small-to-mid herds (less than 1,017 sows), and high productivity (26.5 piglets or more) or ordinary herds (less than 26.5 piglets). A two-level liner mixed-effects model was applied to examine AFM, herd size groups, productivity groups and their interactions for sow life or herd-life performance. No differences were found between either herd size or herd productivity groups for AFM or the number of parity at removal. However, late AFM was associated with decreased removal parity, herd-life days, herd-life piglets born alive and herd-life annualized piglets weaned, as well as with increased sow life days and herd-life nonproductive days ( $P < 0.05$ ). Also, significant two-way interactions between AFM and both herd size and productivity groups were found for longevity, prolificacy, fertility and reproductive efficiency of sows. For example, as AFM increased from 190 to 370 days, sows in large herds decreased herd-life days by 156 days, whereas for sows in small-to-mid herds the decrease was only 42 days. Also, for the same AFM increase, sows in large herds had 5 fewer sow life annualized piglets weaned, whereas for sows in small-to-mid herds this sow reproductive efficiency measure was only decreased by 3.5 piglets. Additionally, for ordinary herds, sows in large herds had more herd-life annualized piglets weaned than those in small-to-mid herds ( $P < 0.05$ ), but no such association was found for high productivity herds ( $P > 0.10$ ). We recommend decreasing the number of late AFM sows in the herd and also recommend improving longevity and lifetime efficiency of individual sows.

**AUTHORS/INSTITUTIONS:** Y. Koketsu, R. Iida, Agriculture, Meiji University, Kawasaki, Kanagawa, JAPAN|C. Piñeiro, PigCHAMP Pro Europa, Segovia, SPAIN|

**CONTROL ID:** 3367324

**TITLE:** Usefulness of self-thinking worksheet for fostering professionalism among biostatisticians

**ABSTRACT BODY:**

**Abstract Body:** The working group of the Biometric Society of Japan developed draft standards of conduct (SOC), which was approved by the executive committee of the Japanese Region in 2013. The English version of the SOC was uploaded to its website <<http://www.biometrics.gr.jp/index.html>>. The SOC offers a framework to encourage individual biostatisticians to establish and hold their own principles (aspirational ethics), since forced guidelines rarely result in full compliance and adherence to ethical conduct.

We developed a program for biostatisticians to think about and make personally meaningful principles in conjunction with the SOC. The program was a two-hour class that comprised lectures and group work, including a case study session to think about the meaning of doing responsible work and a group discussion using the self-development sheet (Mandala chart) to think about the work biostatisticians do. After the class, students were required to fill out a self-thinking worksheet, which consisted of 12 questions asking them about how they feel biostatisticians should act. The worksheet was returned to the student with comments from the instructor.

This program was implemented in the graduate course for biostatisticians at Kyoto University, and 29 students participated. Students provided feedback on the program via questionnaires. All students indicated that the program gave them a chance to think on their own about personally meaningful principles. Some students noted that the program improved their knowledge and skills. In relation to the question on what they thought about prioritizing company interests, many students answered that they should prioritize social interests given the need to gain society's trust.

Most students stated that the self-thinking worksheet was useful, as it allowed them to think about their roles and responsibilities as a professional in society and their own principles. It is difficult to know whether students actually embrace professionalism. However, when one internalizes self-made principles, making excuses for oneself would be akin to self-betrayal. Thus, responsible conduct can be expected. It is necessary for students to dig into own thoughts about whom, and for what they do using tools as self-thinking worksheet. Self-made principles could foster professionalism according to the foundation that is to "do good research and good work."

**AUTHORS/INSTITUTIONS:** K. Sato, Patient Safety, Kyoto University Hospital, Kyoto, Kyoto, JAPAN|T.S. Sato, Biostatistics, Kyoto University School of Public Health, Kyoto, Kyoto, JAPAN|M. Suzuki, Kyoto University Center for iPS Cell Research and Application, Kyoto, Kyoto, JAPAN|

**CONTROL ID:** 3367338

**TITLE:** Multivariate analysis of the physical performance control of amateur boxing athletes

**ABSTRACT BODY:**

**Abstract Body:** Amateur Boxing has improved the training of its athletes. A scientific training was developed with the controls of the tests established for each defined period, for which, Matveet's (1934) sports training theory was used. The objective is to design a scientific training with the use of mathematical-statistical analysis to improve the performance of amateur boxing athletes in correspondence with the requirements of the expected results in the competitions of the INCUFID .The Principal Component Analysis (PCA) was used to select the most important indicators, and later, the Cluster analysis (CA) to group the Athletes according to their similarities. The results showed that the Indicators: Arm flexion 10 s and 60 s, Abdominals 10 s and 60 s, 1000 meters, 30 meters, long jump without impulse, turn in place and bullet launch were necessary for the analysis of the physical performance of Amateur Boxing Athletes, which contributed 97.29% of the total variability.The first component contributed 48.45% of the total variability and synthesized the pure speed, rapid force, resistance strength, explosive leg strength, explosive arm strength (in the first test) and aerobic resistance.The second component contributed 22.07% of the total variability and were represented by abdominal 10 s and 60 s of the first test; abdominals 10 s and 60 seconds of the third test.This component was represented by the ability to force speed and strength endurance. The third and fourth components together contributed 26.77% of the total variability and were represented by left turn, right turn of the first test; left turn, abdominal 60 s and bullet launch of the second test; and bullet launch of the third test. The third component synthesized the coordinative capacity of equilibrium and the fourth component synthesized the resistance force and the explosive force of the arm. The CA showed the four groups with a dissimilarity coefficient of 2.68. The first group was composed of the athletes: Diego, Uriel and Luis; the second group by Ashley; the third group by Cinthya; and the fourth group by Osmani. From the good resistance response, better response at the time of the combat, the positive impact of the designed Scientific Training becomes evident. Although, the first group of Athletes have the best development of physical abilities, all Athletes had the best physical and combative response.

**AUTHORS/INSTITUTIONS:** S. Camelo Avedoy, Department of Biomatematics, , Polytechnic University of the State of Nayarit, Tepic, Nayarit, MEXICO|

**CONTROL ID:** 3367342

**TITLE:** Rare variant imputation with ethnically matched reference panel improves downstream association analyses, revealing new rare variant disease associations

**ABSTRACT BODY:**

**Abstract Body:** Genotype imputation is a standard procedure prior to genome-wide association studies. For common and low-frequency variants, genotype imputation can be performed sufficiently accurately with publicly available and ethnically heterogeneous reference panels like 1000 Genomes Project (1000G). However, the imputation of rare variants has been shown to be significantly more accurate when ethnically matched reference panel is used. Even more, higher genetic similarity between imputation reference panel (IRP) and imputation target samples facilitates the detection of rare and population-specific variants. Notwithstanding, the genome-wide downstream differences in association mapping of rare variants using ethnically mixed and matched IRPs have not been yet comprehensively explored.

We determined and quantified these differences by performing several comparative evaluations of the discovery-driven analysis scenarios using genome-wide array data of ~52,000 Estonians imputed with 1) ethnically matched IRP (N=2,279 Estonians and N=1,856 Finns) and 2) ethnically mixed 1000G IRP (N=2,504). Firstly, gene-based association analyses of rare (minor allele frequency < 1%) coding variants with a set of complex traits (e.g. body mass index, bipolar disorder, Crohn's disease, diabetes) were performed. A generalized mixed model was used to consider sample relatedness and case-control imbalance was accounted for binary traits. The gene-based analysis demonstrated that ethnically matched panel outperformed the 1000G-based imputation, provided a considerable increase in tested genes and significant results. All findings were consequently studied in the UK Biobank (N~400,000) and the FinnGen (N~150,000) cohorts. Secondly, variant-wise analysis replicated several previously reported common variant associations in both imputed datasets, but significant differences were not observed.

These associations provide a solid example of how rare variants can be efficiently analysed to discover novel, potentially functional genetic variants for relevant phenotypes. Furthermore, our work serves as proof of cost-efficient study design, demonstrating that the usage of ethnically matched IRPs can enable substantially improved imputation accuracy of rare variants, facilitating novel high-confidence findings in the rare variant analysis.

**AUTHORS/INSTITUTIONS:** M. Kals, T. Nikopensius, K. Läll, T.T. Sikka, A. Metspalu, T. Esko, R. Mägi, P. Palta, Institute of Genomics, University of Tartu, Tartu, ESTONIA|M. Kals, K. Pärn, S. Ripatti, A. Palotie, P. Palta, Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, FINLAND|T.T. Sikka, Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, ESTONIA|J. Suvisaari, V. Salomaa, National Institute for Health and Welfare, Helsinki, FINLAND|S. Ripatti, A. Palotie, T. Esko, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, UNITED STATES|

**CONTROL ID:** 3367371

**TITLE:** Applications of Meta-analysis for Optimizing Nitrogen Fertilizer in Forage Crops using Non-Linear Mixed Models

**ABSTRACT BODY:**

**Abstract Body:** Fertilization is a key component of an efficient crop system. Economic and environmental constraints imposed by modern agriculture as well as technological advances, such as drip irrigation and fertigation, have changed the focus from applying complete fertilizer formulations to specific nutrient management strategies for supplying crop nutrient needs. The optimal amount of each nutrient needs to be determined using all available information: too little will not be economically feasible, while too much will be environmentally dangerous. Meta-analysis has been used in many areas to synthesize information from independent studies, assessing treatment effects mainly as (standardized) differences between treated and untreated means. In fertilization studies, the optimal amount of a nutrient is called the Crop Nutrient Requirement (CNR), and is computed from a non-linear regression model for predicting the relative crop yield from the fertilizer level. In this paper we study linear and nonlinear mixed effects models to conduct a meta-analysis with a non-linear effect such as this one. We apply these models to summarize nitrogen fertilization requirements in forage crops in the tropics under different soil types, forage species, and crop management.

**AUTHORS/INSTITUTIONS:** R.E. Macchiavelli, D. Sotomayor-Ramirez, Agroenvironmental Sciences, University of Puerto Rico, Mayaguez, PUERTO RICO|A. De Jesus-Soto, Mathematical Sciences, University of Puerto Rico, Mayaguez, PUERTO RICO|

**CONTROL ID:** 3367400

**TITLE:** Constrained Partial Proportional Odds Models for Ordinal Data

**ABSTRACT BODY:**

**Abstract Body:** Partial Proportional Odds Models (PPOMs) have been proposed as a generalisation of Proportional Odds Models (POMs) where the Proportional Odds assumption can be relaxed on one or more of the covariates, corresponding to relaxing the POM assumption of parallel regression lines. However, PPOMS are highly parameterised and this can cause problems. One problem is with convergence, and the full PPOM may not be even be identifiable. Another problem is just as serious; the PPOM can produce negative predicted class probabilities, since as McCullagh and Nelder (1989) noted, non-parallel lines "must eventually intersect".

There are a number of potential solutions to these problems, and in this talk we focus on a reparameterisation via a geometric reformulation of the PPOM, which allows simple imposition of constraints, guaranteeing that estimated and predicted class probabilities will be non-negative. We discuss in detail fitting of these models, comparing an iterative approach with direct estimation. We will also show how this approach enables robust estimation of random effects in PPOMs.

We illustrate the methodology by application to a data set arising from a health survey into environmental attitudes.

**AUTHORS/INSTITUTIONS:** M. Brewer, A. Lorenzo Arribas, Biomathematics and Statistics Scotland, Aberdeen, UNITED KINGDOM|

**CONTROL ID:** 3367413

**TITLE:** Optimizing existing and constructing new sample size recalculation rules

**ABSTRACT BODY:**

**Abstract Body:** Introduction

Adaptive group sequential study designs became one of the key instruments in dealing with planning uncertainties regarding sample size calculations for clinical trials. At interim analyses, adaptive designs allow for an early stopping of the trial or if the trial is continued, provide the possibility of a sample size adaptation for the ongoing trial. There exist multiple ways on how to adapt the sample size [e.g. 1, 2, 3, 4]. Most of them are based on conditional power considerations. So far, guidelines on the application of adaptive study designs [e.g. 5] give no explicit advice on how to choose the sample size adaptation rule.

**Problem**

The decision for a certain sample size recalculation rule is made more difficult by the fact that commonly known recalculation rules often have a high variability in the recalculated sample size or that they do not meet a predefined target power.

**Solution**

Since sample size recalculation rules are functions depending on the observed interim effect, the shape of the respective curve determines the performance of a recalculation design. In this talk, we investigate a smoother increase in the sample size curve as well as bootstrapping of the observed interim effect sizes. These tools can be combined to a new sample size recalculation rule or (partly) applied to existing recalculation rules. We evaluate performance changes by means of Monte Carlo simulations with respect to expected sample size and power, their variations as well as all of these measures combined in a single performance score.

**Discussion**

This approach does not engage with the solution of a mathematical optimization problem. It focusses on easily applicable tools that can be combined or partly adapted to existing sample size recalculation rules.

**References**

- 1) Lehmacher W, Wassmer G: Adaptive sample size calculations in group sequential trials. *Biometrics* 55, 1286-1290 (1999)
- 2) Mehta C, Pocock S: Adaptive increase in sample size, when the results are promising. *Stat Med* 33, 3267-3284 (2011)
- 3) Jennison C, Turnbull B: Adaptive sample size modification in clinical trials: start small then ask for more? *Stat Med* 34, 3793-3810 (2015)
- 4) Hsiao S, Liu L, Mehta C: Optimal promising zone designs. *Biom J*, 1-12 (2018)
- 5) Food and Drug Administration: Adaptive designs for clinical trials of drugs and biologics. Guidance for industry (2019); <https://www.fda.gov/media/78495/download> (accessed Dec 8, 2019)

**AUTHORS/INSTITUTIONS:** C. Herrmann, G. Rauch, Institute of Biometry and Clinical Epidemiology, Charité - University Medicine Berlin, Berlin, GERMANY|C. Herrmann, G. Rauch, BIH, Charité - University Medicine Berlin, Berlin, GERMANY|M. Pilz, M. Kieser, Institute of Medical Biometry and Informatics, University Heidelberg, Heidelberg, GERMANY|

**CONTROL ID:** 3367498

**TITLE:** Semi-parametric survival function estimator under a weak independence assumption of censorship

**ABSTRACT BODY:**

**Abstract Body:** It is a continuing challenge to handling censoring in survival analysis. The most commonly used models are defined as the general random censorship model from the right (GRCM), which assumes that the censoring time is independent of the survival time. However, this assumption is strong and hard to achieve in practice. We propose a new weaker assumption of independence and extend a semi-parameter model developed by Dikta (1998) under this condition. We provide examples and simulation studies to illustrate the validation of the model under our new assumption. The consistency of the extended model is also studied.

**AUTHORS/INSTITUTIONS:** L. Yao, Y. Shao, New York University, New York, New York, UNITED STATES|

**CONTROL ID:** 3367530

**TITLE:** Unified zero-adjusted cure-rate survival models: An application to a sub-Saharan African obstetric dataset

**ABSTRACT BODY:**

**Abstract Body:** Nowadays, survival models accommodating a cure-rate, representing a group of subjects that do not present the event of interest, even after an extended follow-up, are standard. Although it is usually the survival models to assume that the variable time should be greater than zero, there is a diversity of phenomena in which we can find a considerable proportion of individuals with lifetimes equal to zero, which affects the survival curve by deflating its initial value to values smaller than one. For instance, concerning intrapartum care and, notably, the time between hospital admission and vaginal birth, three different patient groups can be observed: a proportion of women with fetal death at admission (representing the zero adjustments), a proportion who undergo to c-section (cure-rate), and a third group, which presents a normal progression, with vaginal delivery (observed labour times). The critical point is that for the patient who arrives at the hospital already having a stillbirth, generally, the time is not registered, leading to the necessity to consider that the time is equal to zero. Statistically speaking, such peculiarity is the so-called zero-adjustment. The survival models, which include these two features, can be referred to as zero-adjusted cure-rate survival models. This class of models is still limited in terms of modeling competitive causes. Therefore, in this paper, we propose a unified version for such survival models, which includes zero-adjusted and cure-rate proportions but also allows for different distributions for the number of latent competitive causes. The application of the proposed modeling is motivated by the peculiarities observed in a real dataset from a sub-Saharan African obstetric study performed by the World Health Organization.

**AUTHORS/INSTITUTIONS:** F. Louzada, P.L. Ramos, ICMC, Universidade de São Paulo, São Carlos, SP, BRAZIL|H.S. Souza, G.D. Perdoná, FMRP, Universidade de São Paulo, Ribeirão Preto, SP, BRAZIL|L. Oyeneyin, Department of Obstetrics and Gynaecology, Mother and Child Hospital, Ondo State, NIGERIA|

**CONTROL ID:** 3367538

**TITLE:** Deep Neural Network for Predicting Neonatal Risk Scores Using Survey Data

**ABSTRACT BODY:**

**Abstract Body:** Recent studies comparing performance of neural networks and Cox PH models in survival prediction have used right censored clinical survival data e.g.- cancer patients' data. However, ability of deep neural networks (DNNs) to predict survival times for events recorded in large scale surveys has remained largely unexplored. In developing countries where reliable follow-up data are generally unavailable, studies of neonatal mortality often rely on survey data. Such data are characterized by high censoring proportions, low number of events-per-variable and lack of information on clinical measurements. Cox PH model, a common tool in survival analysis, may not be able to accurately model the complex, non-linear and convoluted relationships between socio-demographic covariates and neonatal mortality outcome due to the assumption of linearity in the risk function. The objective of this study is to compare the performance of DNNs and Cox PH model in predicting neonatal risk scores using the Bangladesh DHS. Prior to analyzing real data, an extensive simulation study was conducted to compare predictive performances of both approaches under various scenarios encountered in survey data using different types of C-indices. Model parameters and network weights were estimated from training data and applied to prediction in test data. Different regularizations were used with Cox PH and DNN and the latter was tuned for optimized performance through numerous experimentations with hyper-parameters. Simulation results indicated that when the linear risk function was appropriate, DNN performed mostly in parallel with Cox PH and slightly outperformed the latter in scenarios involving heavy censoring and small sample sizes. However, when the non-linear risk function was appropriate, DNN distinctly outperformed Cox PH in all settings thus motivating the use of DNNs. Application to real data revealed that choice of the predictor set had an influence on predictive performance of DNN relative to Cox PH. For example, DNN performed significantly better than Cox PH models regularized by LASSO, ridge or elastic net when the predictor set contained a mixture of socio-demographic, antenatal care and delivery related variables. The study concludes that DNN holds promise in predicting survival times or risk scores using background variables in survey data even in the absence of clinical measurements.

**AUTHORS/INSTITUTIONS:** S. Chowdhury, James P Grant School of Public Health, BRAC University, Dhaka, BANGLADESH|T. Howlader, Institute of statistical Research and Training, University of Dhaka, Dhaka, BANGLADESH|N. Hasan, BRAC University, Dhaka, BANGLADESH|

**CONTROL ID:** 3367564

**TITLE:** A study on Paddy Field Segmentation and Rice Seedling Detection from UAV Images

**ABSTRACT BODY:**

**Abstract Body:** Due to the aging of the population, Taiwan's agricultural manpower has been in a large shortage in recent years. In order to reduce the labor demand of agricultural production, smart farming utilizing big data analysis, Internet of Things (IoT), and sensors is of great interest. Particularly, combining with image data analysis, the inquiries of Unmanned Aerial Vehicles (UAVs) applied in agriculture is rising. It can provide farmers with better management on the process of growing crops. This research proposed an approach to automatically label the rice seedling from images captured by Unmanned Aerial Vehicles (UAVs). There are two parts in the proposed algorithm. First, we extract the paddy field from the UAV images with SLIC superpixel segmentation and classification tool. After that, in the extracted paddy field, we detect the rice seedling with blob detection algorithm. For the extraction of the paddy field, the approach can achieve 91.9% precision and 92.4% recall. For the rice seedling, the approach can achieve 95.5% precision and 71.2% recall. We therefore conclude the proposed methods may be more suitable for image analyses of small-scale paddy fields, which are very common in Taiwan and in some areas in southeast Asia.

**AUTHORS/INSTITUTIONS:** H. Yang, C. Huang, Department of Engineering Science and Ocean Engineering, National Taiwan University, Taoyuan City, TAIWAN|L.D. Liu, Department of Agronomy, National Taiwan University, Taipei, TAIWAN|

**CONTROL ID:** 3367569

**TITLE:** Development and validation of a nomogram based on minor physical anomalies and craniofacial measures for predicting the schizophrenia risk

**ABSTRACT BODY:**

**Abstract Body:** Background: Minor physical anomalies (MPAs) are subtle morphological deficits of the head, face, hands, and feet that are usually determined by the presence of qualitative characteristics and quantitative measurements. Previous studies of embryology have demonstrated that brain and craniofacial morphogenesis are closely related, supporting an association between MPAs and schizophrenia. The MPAs have been found in higher frequencies among individuals with schizophrenia than in healthy controls.

Methods: 508 schizophrenia patients were recruited from five medical institutions in southern Taiwan. For comparison with healthy controls, 281 members of hospital staff and community without a past history of any psychiatric disorder were recruited. We used the patients of three medical institutions and their controls for training and the patients of other two medical institutions and their controls for validation. Univariable and multivariable logistic regression analyses were used to identify the predictive MPAs signature. Finally, a graphic nomogram based on the MPAs signature was developed to predict risk probability of schizophrenia by using rms R package and SAS software.

Results: The minor physical anomalies and craniofacial measures including 41 qualitative and 28 quantitative measurements were identified between schizophrenia and healthy controls. 24 MPAs were found to be associated with risk of schizophrenia ( $P < 0.05$ ) in univariable logistic regression. Among these 24 potential MPAs markers, 11 key predictive MPAs markers were identified by using multivariate analysis. The results showed that the 11-MPAs signature were independent predictive factors. The results of ROC curve analysis showed that the accuracies of development and validation models of 11-MPAs signature in schizophrenia vs. healthy controls were 0.79 and 0.78, respectively. Finally, a graphic nomogram based on the MPAs signature was developed to predict risk probability of schizophrenia.

Conclusions: A MPAs signature associated with onset risk of schizophrenia was constructed and a promising predictive nomogram based on the 11-MPAs signature was developed for prediction of schizophrenia risk in this study.

**AUTHORS/INSTITUTIONS:** S. Lin, H. Tseng, X. Wang, P. Chen, National Cheng Kung University, Tainan, TAIWAN|J. Lin, F. Jang, Chi Mei Medical Center, Tainan, TAIWAN|M. Lu, H. Tan, Jianan psychiatric center, Tainan, TAIWAN|L. Huang, M. Huang, Chia-Yi Branch, Taichung Veterans General Hospital, Tainan, TAIWAN|

**CONTROL ID:** 3367571

**TITLE:** Building blocks for sample size calculations for Soil Microbiome data in agricultural experiments

**ABSTRACT BODY:**

**Abstract Body:** Advances in sequencing have made it feasible to study microbiome composition using 16S marker-gene sequencing in many contexts. However, laboratory, bioinformatic and data analytic efforts needed are substantial enough to warrant careful consideration of the number of samples to be analysed.

Two main types of scientific questions are mostly posed in this type of research:

- 1) are there differences in the overall microbiome composition between treatment groups and
- 2) if so, which species have a higher relative abundance.

For both questions an approach to sample size calculation has been suggested in the literature (Kelly, B. J., et al. (2015).

Bioinformatics 31(15): 2461-2466; La Rosa, P. S., et al. (2012). Plos One 7(12): e52078. Mattiello, F., et al. (2016). Bioinformatics 32(13): 2038-2040). To apply these methods, however, assumptions on measurement variance as well as on expected effects are needed. Since the appearance of these papers, bioinformatic methods have advanced, e.g. sequences are "binned" (aggregated in taxonomic units) using denoising methods rather than statistical clustering methods, while databases used to annotate taxa also expanded. Also, in the cited papers the application context is mainly that of human body sites. Here we present data on both technical and biological variation from agricultural experiments comprising both regular and biological agriculture, using state of the art bioinformatic methods as well as effect sizes of several soil treatments, and apply these to calculate sample sizes for agricultural experiments.

**AUTHORS/INSTITUTIONS:** H.C. Boshuizen, SIM, RIVM, Bilthoven, NETHERLANDS|H.C. Boshuizen, D. te Beest, E. Nijhuis, WUR, Wageningen, NETHERLANDS|

**CONTROL ID:** 3367575

**TITLE:** Quantifying the within plot spatial variation in plant breeding trials

**ABSTRACT BODY:**

**Abstract Body:** Conventional plant breeding partly relies on intuition, skills and judgment by the breeder. Selection based on visual assessment is still the most widely used technique in plant breeding. However, visual grading of phenotypes is time consuming and may suffer from bias. UAV imaging can be used to support the breeder in making objective assessments of traits such as germination and establishment, growth vigour, winter hardiness, flowering and diseases.

When assessing the crop the breeders can choose to look away from a problematic part of a plot or consider this in a rating. The same thing is harder to do when using high throughput UAV imaging.

A quantification of the within plot spatial variation of the crop may have several purposes. It may be used to form a warning indicator, as a high within-plot variation in a plot may be problematic and/or due to outer circumstances, e.g. pests. Alternatively, a measurement of the within-plot variation may be used to adjust for in the statistical analyses. However, most importantly, the within-plot variation can be of direct interest for the breeder who wants uniform plots. The objective of the current work is to suggest an approach to assess the within-plot uniformity based on UAV imaging. To accommodate different breeding contexts, three different crop types will be considered; a large ridge-grown crop, a medium sized crop where the plants are standing isolated in the early growth stages and a small (grain) crop. With a crop-specific approach, UAV imaging shows to be successful in describing within-plot variation.

**AUTHORS/INSTITUTIONS:** S.M. Jensen, J. Svensgaard, S. Christensen, J. Rasmussen, Department of Plant and Environmental Sciences, University of Copenhagen, Taastrup, DENMARK]

**CONTROL ID:** 3367578

**TITLE:** Sex-specific incidence of asthma, rhinitis, and respiratory multimorbidity before and after puberty onset: Individual participant meta-analysis of five MeDALL birth cohorts

**ABSTRACT BODY:**

**Abstract Body:** Introduction Asthma and rhinitis are very common allergic diseases with some still unknown patterns in disease course. To understand the puberty-related sex shift in the prevalence of asthma and rhinitis as single entities and as respiratory multimorbidities, we investigated if there is also a sex-specific and puberty-related pattern of their incidences.

Methods Sex-specific incidence of asthma, rhinitis and respiratory multimorbidity (first occurrence of coexisting asthma and rhinitis) were investigated with harmonized questionnaire-data from 18,451 participants of five prospective observational European birth cohorts within the collaborative MeDALL project. Outcome definitions for IgE- and non-IgE-associated asthma and rhinitis were based on questionnaires and specific antibodies (IgE) against common allergens in serum. For each outcome, we performed a one stage individual participant data (IPD) meta-analysis in which all data is analyzed simultaneously, accounting for the cohort clusters. Proportional hazard models including puberty as a time dependent covariable with the average partial likelihood method for handling ties in the event times were used for analysing the data. The focus was on the interaction puberty\*sex as an indicator of sex-specific changes in allergy incidence in participants before versus those in or after puberty.

Results Girls had a lower risk of incident asthma (adjusted Hazard Ratio 0.67, 95%-CI 0.61-0.74), rhinitis (0.73, 0.69-0.78) and respiratory multimorbidity (0.58, 0.51-0.66) before puberty compared to boys. After puberty onset, these incidences became more balanced across the sexes (asthma 0.84, 0.64-1.10; rhinitis 0.90, 0.80- 1.02; respiratory multimorbidity 0.84, 0.63-1.13). The incidence sex shift was slightly more distinct for non-IgE associated respiratory diseases (asthma 0.74, 0.63- 0.87 before versus 1.23, 0.75-2.00 after puberty onset; rhinitis 0.88, 0.79-0.98 versus 1.20, 0.98-1.47; respiratory multimorbidity 0.66, 0.49-0.88 versus 0.96, 0.54-1.71) than for IgE-associated respiratory diseases.

Discussion We found an incidence 'sex shift' in chronic respiratory diseases from a male predominance before puberty to a more sex-balanced incidence after puberty onset that may partly explain the previously reported sex shift in prevalence.

**AUTHORS/INSTITUTIONS:** T. Keller, Institute for Biometry and clinical Epidemiology, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Berlin, GERMANY|C. Hohmann, S. Roll, T. Keil, Institute of Social Medicine, Epidemiology and Health Economics, Charité – Universitätsmedizin, Berlin, GERMANY|A. Wijga, National Institute of Public Health and the Environment, Bilthoven, NETHERLANDS|U. Gehring, Department of Pulmonology, University Medical Center Groningen, University of Groningen, Groningen, NETHERLANDS|M. Standl, J. Heinrich, Institute of Epidemiology, Helmholtz Zentrum München – German Research Centre for Environmental Health, Neuherberg, GERMANY|I. Kull, Department of Clinical Science and Education, Soedersjukhuset, Karolinska Institutet, Stockholm, SWEDEN|A. Bergström, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, SWEDEN|I. Lehmann, Department of Environmental Immunology/Core Facility Studies, Helmholtz Centre for Environmental Research– UFZ, Leipzig, GERMANY|I. Lehmann, Environmental Epigenetics and Lung Research Charité – Universitätsmedizin Berlin, Berlin, GERMANY|A. von Berg, Research Institute, Department of Pediatrics, Marien-Hospital Wesel, Wesel, GERMANY|S. Lau, U. Wahn, Department of Paediatric Pneumology & Immunology, Charité – Universitätsmedizin Berlin, Berlin, Germany, Berlin, GERMANY|D. Maier, Biomax Informatics AG, Munich, GERMANY|J. Antó, Centre for Research in Environmental Epidemiology (CREAL), ISGlobal, Barcelona, Spain, Barcelona, SPAIN|J. Antó, Hospital del Mar Research Institute (IMIM), Barcelona, SPAIN|J. Bousquet, University Hospital Montpellier, Montpellier, FRANCE|J. Antó, J. Bousquet, Universitat Pompeu Fabra (UPF), Barcelona, SPAIN|J. Smit, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, NETHERLANDS|T. Keil, Institute for Clinical Epidemiology and Biometry, University of Würzburg, Würzburg, GERMANY|

**CONTROL ID:** 3367585

**TITLE:** Checking the Goodness-of-Fit for Right-Censored data With the R package GofCens

**ABSTRACT BODY:**

**Abstract Body:** Goodness-of-fit techniques are an important tool to test the validity of parametric models and to provide indications that the modeling assumptions are reasonable. No general asymptotic optimality theory exists for this very difficult problem (Lehmann 2005); in fact, any test can achieve high asymptotic power or perform uniformly well against local or contiguous alternatives when the family of possible alternatives is large (Jansen 2000).

Kolmogorov-Smirnov and chi-squared goodness-of-fit tests encompass the most used analytical tests. However, because of the lack of good power of these tests, graphical techniques to assess the validity of distributional assumptions and to check goodness-of-fit should always tag along with the tests.

Goodness-of-fit tests have been developed for complete data and based either on the empirical distribution function or on chi-squared type tests. Preliminary extensions to account for right-censored data were proposed by Barr and Davidson (1973) for life-testing applications, who modified Kolmogorov-Smirnov statistics for censored or truncated data. Koziol and Green (1976) developed Cramer-von Mises type statistics based on the product-limit empirical distribution function, when the data are subject to random censorship. Mihalko and Moore (1980) were pioneers proposing an extension of chi-squared tests to censored data, based on the introduction of random cells where the boundaries of the intervals depend on some summaries of the data.

We present the R package GofCens for random right-censored data which provides for nine different parametric models: i) Graphical plots such as P-P plot, Q-Q plot, Stabilized probability, Empirically rescaled plot, Cumulative Hazard Plot to be used as exploratory techniques; ii) Tests based on the empirical distribution function, in particular, Kolmogorov-Smirnov and Crámer-von Mises tests adapted to right-censored data; and iii) Chi-squared type tests based on the squared difference between observed and expected counts by means of random cells and that can be used either for continuous or discrete data. We present the main functions of GofCens and illustrate them with real data examples.

**AUTHORS/INSTITUTIONS:** G. Gomez Melis, Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Barcelona, SPAIN|M. Besalú, Genetics, Microbiology and Statistics, Universitat de Barcelona, Barcelona, SPAIN|K. Langohr, Universitat Politècnica de Catalunya, Barcelona, SPAIN|

**CONTROL ID:** 3367592

**TITLE:** Shared frailty model applied to incidences for lameness removal of sows and their reproductive performance

**ABSTRACT BODY:**

**Abstract Body:** Lameness is the one of the major reasons for sow removal, and increased lameness occurrences can decrease reproductive efficiency and increase a welfare concern. The objectives of this study were to estimate the incidence rate of lameness removal and to characterize the reproductive performance and longevity of sows that were removed due to lameness. Shared frailty models were applied to a cohort data of 165,918 sows in 148 Spanish herds taking account of between-herd variability. The Wilcoxon rank sum test was used to compare the performance between sows removed due to lameness and their controls in one-to-two matched case-control datasets. The incidence rate for the removal was 19.9 cases per 1000 sow-years. Also, 72.6% of 6784 lameness removal occurred in farrowed sows, whereas the remaining happened in serviced sows. In serviced sows, factors associated with lameness removal were re-service and the 4-5 weeks after service ( $P < 0.01$ ). For example, the removal incidence was 2.12 times higher in re-serviced sows than in first serviced sows ( $P < 0.01$ ), and was also 1.57-4.87 times higher at the 4-5 weeks after service than during the first 4 weeks ( $P < 0.01$ ). However, the lameness removal was not significantly associated with parity ( $P = 0.07$ ). Meanwhile, in farrowed sows, a higher incidence of lameness removal was associated with higher parity and the 4-8 weeks after farrowing ( $P < 0.01$ ). The incidence was 1.32-1.73 times higher in parity 4-5 than in parity 1-2, and was 2.50-38.94 times higher at 4-8 weeks after farrowing than during the first 4 weeks ( $P < 0.01$ ). Lastly, in the case-control datasets, sows removed due to lameness had higher weaning-to-first mating interval (means: 6.5 vs. 5.8 days), lower number of piglets born alive (11.7 vs. 12.5 piglets) and lower parity at removal than the control sows (3.4 vs. 4.9;  $P < 0.01$ ). However, no difference was found in age at first service between the case and control groups ( $P \geq 0.53$ ). In conclusion, sows that were removed due to lameness could have had subclinical feet and legs problems in a previous stage in their reproductive cycle, and consequently might have had lower farrowing performance than the control sows. We recommend checking subclinical lameness at each stage of the reproductive cycle and making a quick decision to cull a sow at risk in order to decrease the welfare concern.

**AUTHORS/INSTITUTIONS:** R. Iida, Y. Koketsu, School of Agriculture, Meiji University, Kawasaki, JAPAN|C. Piñeiro, Department of Data Management and Analysis, PigCHAMP Pro Europa S.L., Segovia, SPAIN|

**CONTROL ID:** 3367610

**TITLE:** Sparse genetic prediction modeling incorporating gene-environment interactions

**ABSTRACT BODY:**

**Abstract Body:** We propose a genome-wide genetic prediction modeling that can include not only marginal gene effects but also gene-environment interaction effects. The proposed approach is a straightforward extension of our previously developed regression-based method, STMGP (smooth-threshold multivariate genetic prediction; Ueki and Tamiya 2016), in which the genome-wide test statistics for gene-environment interaction analysis are used to weigh the corresponding genetic variants. Our model enables sparse modeling in the sense that irrelevant predictors (either genetic variants or gene-environment combinations) are removed from the model, and also incorporates multiple environment variables simultaneously. We carry out simulation studies to evaluate the proposed methodology in comparison with other modeling approaches. We also provide a real genome-wide data application using Alzheimer's Disease Neuroimaging Initiative (ADNI) data to evaluate our prediction model in real-world GWAS (genome-wide association study) data.

**AUTHORS/INSTITUTIONS:** M. Ueki, RIKEN Center for Advanced Intelligence Project, Tokyo, JAPAN|

**CONTROL ID:** 3367619

**TITLE:** Interval Estimates of the Probability of Toxicity at the Recommended Dose Following a Phase I Clinical Trial

**ABSTRACT BODY:**

**Abstract Body:** Phase I clinical trials in cancer are conducted to reach a safe and effective dose that can be used in phase II. This study concerns the first model-based dose-finding design, the continual reassessment method (CRM). It was proposed to meet ethical considerations by avoiding treating a lot of patients at low dose levels, which might be ineffective, for a longer time as in most popular up-and-down designs. This method starts the trial with a level that is above the lowest one and below the highest level, and believed to be close to the target toxicity level, to be able to escalate and de-escalate the doses in the trial. Generally, the CRM is based on a mathematical model for the relationship between the probability of toxicity and the dose. It sequentially uses the previous information about a patient's toxic response to estimate the dose that is close to the target dose that needs to be assigned to the next patient, mostly using Bayesian methodology.

At the end of the trial, estimates of the probability of toxicity and a confidence interval can be derived using either Bayesian or maximum likelihood approaches. Such a confidence interval for the probability of toxicity can be used to support the decision about identifying the maximum tolerated dose (MTD). Therefore, this study aims to evaluate the confidence interval for the toxicity probability at the MTD by assessing its width and coverage probability. Due to the small number of patients in phase I trials, a wide confidence interval for the probability of a toxic response is obtained. However, it has been shown that the coverage of the confidence interval is close to the nominal value when using the Cornish-Fisher correction for sample sizes as small as 12.

A logistic function is assumed here for the dose-toxicity model and a broad range of probabilities of toxicity are considered. Furthermore, different approximation methods are applied to determine whether a narrower interval for the probability of toxicity can be found while maintaining a coverage close to the nominal value. The simulation results show that the coverage is close to the nominal value in most cases considered. So it can be concluded that using a confidence interval for the probability of toxicity is accurate enough to support the decision in recommending a dose as the MTD.

**AUTHORS/INSTITUTIONS:** A. Althobety, Statistics, King Abdulaziz University, Jeddah, SAUDI ARABIA|A. Althobety, School of Mathematical Sciences, Queen Mary, University of London, London, UNITED KINGDOM|

**CONTROL ID:** 3367620

**TITLE:** Conditional restricted mean survival time – an intuitive and easy-to-interpret measure

**ABSTRACT BODY:**

**Abstract Body:** Compared with the hazard ratio (HR), restricted mean survival time (RMST) does not rely on the proportional hazards assumption and is clinically interpretable, thus, recently it has received more and more attention. RMST, however, which estimates the average survival time from the start of follow-up to a given time horizon ( $w$ ), is not very informative for a patient who has already survived several years after initial diagnosis or treatment. In other words, it is not directly capable of dynamic analysis or prediction. To this end, RMST is extended to the conditional restricted mean survival time (cRMST), which is the restricted mean survival time of a patient for further  $w$  years, given that he/she has already survived  $s$  numbers of years. In this article, we introduce the estimation of cRMST based on pseudo-observations, hypothesis testing for the cRMST difference between two groups, extended regression analysis, as well as the dynamic prediction model. The Monte-Carlo simulation results show that the proposed hypothesis test is well specified under the null hypothesis (that is, the type I errors are all close to 0.05 as expected in different scenarios) and has good power properties under the alternative hypothesis. For illustration purpose, several examples are provided and the analysis results show that cRMST, which depends on the prediction time  $s$ , can dynamically reflect how prognosis changes over time and offer more scientific prognostic information for patients who are fortunate enough to have survived initial  $s$  years.

**AUTHORS/INSTITUTIONS:** Z. Yang, Z. Chen, Department of Biostatistics, Southern Medical University, Guangzhou, CHINA|Y. Hou, Department of Statistics, Jinan University, Guangzhou, CHINA|

**CONTROL ID:** 3367621

**TITLE:** Spatial-temporal Distribution of Tuberculosis and HIV in Ethiopia: Generalized Linear Mixed Model Application

**ABSTRACT BODY:**

**Abstract Body:** Tuberculosis (TB) has claimed many lives and it continues to be a global threat in the coming decades, especially in developing countries like Ethiopia. Being preventable, TB remains still challenging and second leading cause of death next to HIV in Ethiopia. The general objective of this study is to assess spatio-temporal distribution of TB, detect TB/HIV clustering and identify hot-spot areas. Five years aggregated data of TB and HIV cases of 804 districts were used for the analyses. Global and local spatial test statistics, namely Moran's I and Geary's were used for measuring spatial autocorrelation. Exploratory spatial analysis shows clustering in TB distribution. Local indicator of spatial autocorrelation (LISA) cluster map shows hot spot located in North West and Central part of the country while cold spot areas covering most of Southern part. Furthermore, there are between districts and within district variations in TB distributions over time. Therefore, the generalized linear mixed effects models (GLMMs), which incorporate random and fixed effect were used to model the number of TB cases. Among the GLMMs considered, model with quadratic random effect fitted the data well. The results also show decline in rate of change over five years period; however, the results confirm the presence of spatial autocorrelation between the number of TB and HIV cases in Ethiopia. The identified hot-spot areas require special attention and immediate intervention by government and concerned bodies.

**AUTHORS/INSTITUTIONS:** L.L. Gemechu, L. Debusho, Statistics, University of South Africa, Johannesburg, Roodeport, Johannesburg, SOUTH AFRICA]

**CONTROL ID:** 3367624

**TITLE:** Center heterogeneity in treatment effect in a randomized trial on childhood acute lymphoblastic leukemia (AIEOP ALL 2000 study)

**ABSTRACT BODY:**

**Abstract Body:** Important between-center differences in outcome of children enrolled in multicenter randomized trials are of importance in the approach to data analysis and in the interpretation of trial results. Our aim here is to investigate with different methodologies the variability of treatment effect with application in the context of a large paediatric randomized trial on Acute Lymphoblastic Leukemia (ALL). We used individual data from the AIEOP-ALL-2000 trial which enrolled 1999 children between 2000 and 2006 and included a randomized question on the administration of the standard steroid therapy with prednisone or the experimental therapy with dexamethasone as part of a 4-drug induction therapy. Randomization was 1:1, stratified by center size, and performed in 41 centers in Italy. The primary endpoint was Event Free Survival (EFS), where events were nonresponse, relapse, secondary neoplasm or death from any cause, while secondary endpoint was Overall Survival (OS). Follow-up was updated as of January 2014.

After a standard exploratory analysis on EFS and OS estimated curves, a model based analysis was applied where the center specific hazards ratios (HR) of dexamethasone vs prednisone were estimated from a Cox Proportional Hazards model adjusting for patients' characteristics and including a center-specific random effect (i.e. frailty) to account for the clustered nature of data. Gender, age at diagnosis, white blood count at diagnosis, immunophenotype, genetic features, response to Prednisone at day +8, Minimal Residual Disease at the end of induction and consolidation were used for adjustment. The median hazard ratio (MHR, Austin et al. 2017) was also calculated as a measure of variability between centers.

A large variability in HRs was shown across centers and further investigations should clarify the nature of the unexplained heterogeneity, such as residual confounding and structural between-center differences. This might be important for the design and analysis of future multicenter randomized protocols.

**AUTHORS/INSTITUTIONS:** S. GALIMBERTI, F. Graziano, D. Silvestri, V. Conter, M. Valsecchi, Department of Medicine and Surgery, University of Milano-Bicocca, Monza, ITALY|D. Silvestri, V. Conter, Department of Pediatrics, Pediatric Hematology-Oncology Unit, MBBM Foundation/ASST Monza, Monza, ITALY|

**CONTROL ID:** 3367626

**TITLE:**

A review of factor analysis and related methods for handling longitudinal data

**ABSTRACT BODY:**

**Abstract Body:** In recent years, there has been a rapid increase in the application of factor analysis for longitudinal data. Since longitudinal data are characterized by a complex pattern of variability, different factor models and approaches have been developed to handle such data structure. However, many researchers are not familiar with statistical methods related to this area and accordingly, they are not able to identify and implement the model applicable to the context of their own research. Therefore, they often apply methods provided by popular software packages which may lead to wrong conclusions if the applied method is incorrect with respect to the context or data. The present work provides an overview of factor analysis and related methods for the analysis of longitudinal data including multi-group factor analysis, multilevel factor analysis and dynamic factor analysis. The aim of this review is to provide an improved understanding of these methods and then suggest recommendations that can help for the choice of their application.

**AUTHORS/INSTITUTIONS:** B. TRAORE, Mathematics, University of Lagos, Lagos, Lagos, NIGERIA|

**CONTROL ID:** 3367628

**TITLE:** Perspectives on external validation and generalizability of multivariable prediction models

**ABSTRACT BODY:**

**Abstract Body:** Clinical literature is replete with multivariable prediction models (MPMs) developed to diagnose diseases or to predict their course. Interestingly, the majority of these MPMs are rarely used in clinical practice, which is, among others, due to lack of confidence in their performance beyond the studies in which they are developed. Hence, assessment of generalizability of an MPM is essential for ensuring its acceptability in clinical practice. Among others, the assessment entails model validation for deciding if a model is transportable or merely reproducible. Consequently, issues related to model validation attract considerable attention in medical statistics.

Validation of an MPM is a multifaceted process for assessing the performance of the developed model in a specific population. The MPM can be a regression model of any kind, an artificial neural network, a classification tree, or the like. It is crucial that the MPM is considered given and fixed during the validation process. Sometimes an MPM is altered or updated at the end of a validation exercise; this, however, has to be considered equivalent to the development of a new MPM, which again would need proper validation.

Model validation can be performed using internal or external validation approaches. We focus on external validation of an MPM; we identify some lack of clarity in the main concepts and a frequent loose use of key terms. We suggest a clarification and harmonization of the key features; these are credibility, relatedness, performance, clinical utility, and generalizability (transportability versus reproducibility). A further term that is often used without reliable definition is the so-called case-mix, which refers to patient characteristics in a data set.

The aforementioned concepts and terms will be addressed and illustrated using external validation of an MPM for pulmonary arterial hypertension.

**AUTHORS/INSTITUTIONS:** H. Heinzl, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, AUSTRIA|H. Chadha-Boreham, Clinical Biostatistics Consultancy, Dijon, FRANCE|

**CONTROL ID:** 3367660

**TITLE:** Sensitivity of regression calibration to misspecification in the outcome or error model

**ABSTRACT BODY:**

**Abstract Body:** It is well known that many exposures commonly of interest in epidemiologic studies are prone to measurement error (ME), which can distort the exposure-outcome relationships. This is particularly true in nutrition epidemiology given that self-reported intakes tend to be subject to both systematic and random (unbiased) measurement error. For some nutrients, recovery biomarkers that have unbiased (classical) measurement error exist. Regression calibration (RC) has been shown to correct bias in outcome regression models induced by exposure ME. While it is well-understood that misspecifying the regression models leads to biased parameter estimates, how robust RC is to misspecification has been understudied.

The study purpose was to investigate how misspecification of either the RC model or outcome model affects the performance of RC. We assume the main exposure of interest is subject to both systematic and random errors (SME), while on at least the subset there is an additional measure with unbiased (classical) measurement error (CME). Extensive simulations under SME and CME were performed that examined the effects of leaving a covariate out of 1) the RC model 2) the outcome model or 3) both under varying scenarios of correlation between the left-out covariate and other predictors.

Estimates of the outcome regression parameters were greatly biased when a covariate was omitted in RC models and correlation structures between each covariate determined the bias direction (either attenuation or inflation). Specifically, when a key covariate was omitted in the RC model but kept it in the outcome model, estimates of the covariate with CME and other covariate were greatly attenuated under high correlation, while the coefficient of the key covariate was largely inflated. The similar pattern was observed in data with SME with greater magnitude of changes in coefficients. Perhaps unexpectedly, when a covariate was omitted in the RC model under SME with no correlation between covariates, there was appreciable bias in all outcome-model coefficients. Misspecification in the outcome model led to similar patterns.

To be assured of unbiased results, both the outcome and RC models need to be correctly specified, which may be difficult to do reliably in practice. Sensitivity analyses are an important tool to investigate the robustness of study results to varying levels of misspecification.

**AUTHORS/INSTITUTIONS:** E. Park, P. Shaw, Graduate Group in Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES|D. Sotres-Alvarez, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, UNITED STATES|P. Gustafson, Department of Statistics, University of British Columbia, Vancouver, British Columbia, CANADA|

**CONTROL ID:** 3367666

**TITLE:** Bayesian Semi-Parametric Linear Mixed Model Using Smoothing Spline: Application to longitudinally measured Fasting Blood Sugar level data

**ABSTRACT BODY:**

**Abstract Body:** In the analysis of longitudinal data, the mean profile is often estimated by parametric linear mixed effects model. However, individual and mean profile plots of the data sometimes might exhibit a nonlinear form and imposing parametric models may be too restrictive and yield unsatisfactory results. We propose a Bayesian semi-parametric mixed model, in particular using spline smoothing to analyze efficiently a longitudinal measured fasting blood sugar level of adult diabetic patients at the diabetic clinic of Jimma University Specialized Hospital in Ethiopia, accounting for correlation between observations through random effects.

The proposed model has improved the efficiency of parameter estimates and captured well the longitudinal change of the fasting blood sugar level data of diabetic patients. The study revealed that the rate of change in FBS level in diabetic patients, due to the clinic interventions, does not continue as a steady pace but changes with time and weight of patients. Therefore, the individual profile curve can be used to follow patient's specific FBS level trends over time and hence might help to monitor the individual-level trends of patients over time.

Keywords: Diabetes mellitus, fasting blood sugar level, Bayesian semi-parametric linear mixed model

**AUTHORS/INSTITUTIONS:** T. Aniley, L. Debusho, T.A. Diriba, Statistics, University of South Africa, Roodepoort, Gauteng, SOUTH AFRICA|

**CONTROL ID:** 3367677

**TITLE:** Statistical issue in tumor measurement data of cancer trial

**ABSTRACT BODY:**

**Abstract Body:** Background: The Response Evaluation Criteria for Solid Tumors (RECIST) has been a standard tool to assess treatment effect in oncology clinical trials and has helped advance cancer treatment. However, studies have raised issues regarding RECIST being applied for incorrect determination of response, which resulted in premature termination of therapy and imprecise efficacy. RECIST measures a set of lesions ('target lesions') over time and takes the sum of all lesion sizes for clinical decision-making. In this study, we evaluate whether the RECIST aggregated sum metric is an optimal assessment tool for treatment efficacy. We use our institute clinical trial data to examine the issue and provide our suggestions.

Methods and Results: Tumor measurement data collected in target lesions during a clinical trial often shows variation of tumor growth among lesions in a patient. Some lesions have fast tumor growth, but some have slow growth, and some even have tumor reduction. Two types of lesion heterogeneity are observed from our trial data: (i) variation with opposite trend of tumor growth in different organ sites or in an organ site: some lesions with fast growth rate but other lesions with tumor reduction; (ii) variation in growth scale with same trend in different organ sites or in an organ site: some lesions with fast growth (regression) but other lesions with slow growth (reduction). We find a significant patients (>30%) with lesion heterogeneity in our cancer clinical trial data. Literature also report about 20% patients with opposite trend of tumor growth. These data indicate lesion heterogeneity is a not rare event.

Given abundance of lesion heterogeneity, we examine how the aggregated sum metric of all lesion sizes affect RECIST performance. We compare % change (from baseline) between the sum and individual lesion and find high degree of deviation. Such difference suggests the RECIST's sum does not well capture lesion tumor growth unless lesion heterogeneity is taken into consideration. The issue becomes more serious when lesions have opposite trend of tumor growth. In this case, the assessment of tumor growth could give discrepant results between RECIST and individual lesion and could lead to controversial clinical decision. Nonlinear mixed effect model could be a potential tool to fully utilize all lesion data to provide overall assessment of treatment effect and individual lesion variability.

**AUTHORS/INSTITUTIONS:** D. chen, Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, Florida, UNITED STATES|W. chan, The University of Texas Health Science Center at Houston, Houston, Texas, UNITED STATES|B. Creelan, T. Rose, Moffitt Cancer Center, Tampa, Florida, UNITED STATES|

**CONTROL ID:** 3367708

**TITLE:** Regulation of medical devices: detecting adverse effects and the cost of ambiguous diagnostic definitions

**ABSTRACT BODY:**

**Abstract Body:** Biostatisticians have made important contributions to the regulation of pharmaceutical treatments, before and after licencing. As scientists, we can contribute to regulation by alerting agencies or industry of possible risks, and evaluate the scale of the risk. Some statisticians participate in setting regulatory standards, or assessing liability and the magnitude of compensation.

Implanted medical devices provide new challenges, particularly for discovering, defining and monitoring adverse reactions, or when a company fails. Decision-making once an adverse event is confirmed might lead to operations. As operations are inherently risky, the balance of risks has to be considered. Operations to remove well-functioning devices in order to limit a company's liability, because the company has stopped trading might not be a patient's choice. Regulators in different jurisdictions might wish to consider alternative approaches to discontinued devices which patients find beneficial.

Multibillion dollar lawsuits have been brought in relation to metal-on-metal hip replacements. In 2012, after initial reports, the UK regulator recommended screening people with metal-on-metal hip replacements for 'Adverse reaction to metal debris (ARMD)' using measurements of cobalt or chromium blood levels. Blood tests were used as part of diagnostic procedures. However, even the definition of ARMD was not agreed, so there could be no confidence in the quality of diagnoses.

We document the introduction of the expression ARMD, its development and the range of definitions. We investigate how hips which 'failed' due to ARMD were identified in practice by reviewing the research literature. The quality of reporting of studies of diagnostic accuracy is summarised. The implications for estimates of accuracy of alternative definitions is discussed.

Assessment of blood metal ions as diagnostic tests for ARMD, following the full systematic review process, is premature. A reliable and valid definition of ARMD is a prerequisite. We recommend further debate about processes for identifying and responding to long-term adverse consequences of implanting medical devices.

**AUTHORS/INSTITUTIONS:** J.L. Hutton, L. Nichols, Statistics, University of Warwick, Coventry, UNITED KINGDOM|

**CONTROL ID:** 3367712

**TITLE:** A logistic mixed effects model to robustly test for rare variant association with common controls

**ABSTRACT BODY:**

**Abstract Body:** Genetic summary data from public resources, such as the genome Aggregation Database (gnomAD), can be used as common controls (i.e. convenience controls) in case-control analysis. Using common controls can increase sample size and subsequently power of rare variant aggregation tests. Further, genetic summary data, compared to individual level data, often has fewer barriers to access promoting open science and the broad use of valuable resources. However, summarizing individual-level data can mask heterogeneity within and between samples, such as population structure or differences in sequencing technology and variant calling pipelines, resulting in increased bias and type I error for traditional statistical tests. Methods, such as iECAT-O and ProxECAT, use allele frequencies from common controls while maintaining the appropriate type I error. However, these methods cannot adjust for covariates or include cases or controls from multiple sources.

Here, we present a method to address these limitations by changing the observational unit (the unit by which we gather and analyze data) from the individual to the genetic alternate allele. By doing so, we regain information on the observational unit level and are able to harness traditional statistical frameworks such as a logistic mixed effects model. Our model enables the inclusion of alternate allele-level and sample-level covariates. Alternate allele-level covariates include the proportion of alternate variant reads or depth of coverage. Sample-level covariates enable the inclusion of both internal and external controls. Population structure can be modeled as fixed or random effects. We evaluate our method over a wide variety of simulation scenarios and in real data. We find that our method maintains the expected type I error rate even in the presence of batch effects between cases and common controls. For moderately sized case samples (e.g.  $N=1000$ ), we find the power increases as the number of common controls increases to that currently available in control databases.

In summary, we enable the use of common control data in a general model framework that can incorporate covariates and samples from multiple sources. Our method has the ability to increase the sample size and subsequently power to detect rare genetic variant associations while controlling for differences between cases and controls.

**AUTHORS/INSTITUTIONS:** A. Hendricks, Mathematical and Statistical Sciences, University of Colorado Denver, Denver, Colorado, UNITED STATES|A. Hendricks, Biostatistics and Informatics, Colorado School of Public Health, Aurora, Colorado, UNITED STATES|M.M. Null, University of Colorado Denver, Aurora, Colorado, UNITED STATES|J. Dupuis, Biostatistics, Boston University School of Public Health, Boston, Massachusetts, UNITED STATES|

**CONTROL ID:** 3367714

**TITLE:** Adaptively weighted Fisher's meta analysis method and applications in omics studies

**ABSTRACT BODY:**

**Abstract Body:** Meta-analysis methods have been widely used to combine results from multiple clinical or genomic studies to increase statistical powers and ensure robust and accurate conclusions. The adaptively weighted Fisher's method (AW-Fisher), initially developed for omics applications but applicable for general meta-analysis, is an effective approach to combine P-values from K independent studies and to provide better biological interpretability by characterizing which studies contribute to the meta-analysis. Currently, AW-Fisher suffers from the lack of fast P-value computation and variability estimate of AW weights. In this paper, we develop an importance sampling scheme with spline interpolation to increase the accuracy and speed of the P-value calculation. We also apply bootstrapping to construct a variability index for the AW-Fisher weight estimator and a co-membership matrix to categorize (cluster) differentially expressed genes based on their meta-patterns for intuitive biological investigations.

**AUTHORS/INSTITUTIONS:** S. Tang, Roche Molecular System, Inc, Pleasanton, California, UNITED STATES|Z. Huo, Biostatistics, University of Florida, Gainesville, Florida, UNITED STATES|Y. Park, G. Tseng, Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, UNITED STATES|

**CONTROL ID:** 3367718

**TITLE:** New proposal of indicators of soil quality: a nonlinear multivariate approach.

**ABSTRACT BODY:**

**Abstract Body:** The top layer of the earth's crust, the soil, is a variable and complex medium that plays an important role as interface between earth, air and water, and performs many vital functions: habitat and gene pool, medium of food and other biomass production, container of carbon, provider of raw materials that simultaneously serves as a platform for human activities, heritage and landscape. Assessment of soil quality (SQ) has gained widespread interest as indirect way to evaluate, protect and preserve this medium that can be considered as a non-renewable resource, given that their formation is a really very slow process.

SQ is an abstract concept that cannot be measured directly, so conceptually will be controversial. Nevertheless, many different SQ evaluation methods have been developed in the last times: from qualitative approaches to quantitative indexes using mathematic models or statistical methods. SQ indices (SQI) obtained by integrating mainly physicochemical soil property indicators through the use of traditional multivariate methods, are the most commonly used. Mostly of these SQI have been developed for quantitative soil indicators but they don't consider relevant qualitative aspects such the morphology, type and conditions of the soils, vegetation and land use; but neither socioeconomic, geographic and climatic issues being inherently linked to pedogenic processes. Assessment of SQI using only the quantitative surface soil properties provides an incomplete information on the subject.

We promote the application of the Gifi system of nonlinear multivariate approach (Gifi, 1990) in order to achieve a more comprehensive SQ index. By using optimal scaling methods, we can analyze all the different variables measured in different scales in two ways: to determinate among them the most informative; furthermore, to obtain a low-dimensional solution on which to build most suitable SQI. This novel approach in the construction of an SQI, is particularly relevant given that the use of a limited number of soil indicators (but both quantitative and qualitative) reduces the analysis cost and therefore, allows increasing the sampling density for large-scale evaluations.

A practical example of the construction of an SQI with real data, will be presented.

**AUTHORS/INSTITUTIONS:** C. Avila-Zarza, Statistics, Faculty of Agricultural and Environmental Sciences. University of Salamanca, Salamanca, SPAIN|F. Santos-Frances, Soil Sciences, Faculty of Agricultural and Environmental Sciences. University of Salamanca, Salamanca, SPAIN|

**CONTROL ID:** 3367723

**TITLE:** The prevalence of Hepatitis B surface antigen among pregnant women in the Limbe and Muyuka Health Districts of the South West region of Cameroon: a three-year retrospective study

**ABSTRACT BODY:**

**Abstract Body:** Introduction: Hepatitis B infection is caused by the Hepatitis B Virus (HBV). HBV infection is a major health problem worldwide owing to its high prevalence and significant morbidity and mortality. It is transmitted through sexual intercourse, by exchange of saliva during kissing and also to newborns of infected mothers. There are about 2 billion people living with HBV worldwide and over 360 million chronic carriers. In the global burden of diseases 2010, 786,000 deaths were attributed to HBV. Studies in Cameroon, reported the prevalence of HBV as high as 10.1% and 12% among blood donors in hospital blood banks. This study aims at determining the prevalence of Hepatitis B surface antigen (HBsAg) among pregnant women; assess the knowledge and practice of pregnant women and health care workers in the antenatal clinic (ANC) and maternity units on HBV infection and transmission.

Methods: ANC registers were exploited from the health centers for a period of three years (2014-2016) in order to determine the prevalence of HBV infection. About 270 pregnant women attending ANC and 31 health care workers were selected by exhaustive sampling. Knowledge and practice of the participants on HBV prevention and transmission was assessed using a structured questionnaire.

Results: The prevalence of HBV in the Limbe Health District (LHD) and Muyuka Health District (MHD) were 5.7% and 7.5% respectively. Pregnant women in the LHD demonstrated good knowledge but adopted poor practices, whereas in the MHD, pregnant women demonstrated poor knowledge and adopted poor practices regarding the mode of transmission and prevention of HBV infection. There was a significant association between the prevalence of HBsAg and marital status ( $p = 0.000$ ) in the LHD and age ( $p = 0.022$ ) in the MHD.

Conclusion: This study indicated a high prevalence of HBV among pregnant women in the LHD and MHD. There was a significant association between the knowledge and practice of pregnant women and health care workers on Hepatitis B prevention in the Muyuka Health District ( $P = 0.0006$ ).

**AUTHORS/INSTITUTIONS:** B.M. Yankam, Statistics, University of Nigeria, Nsukka, Nsukka, Enugu, NIGERIA|B.M. Yankam, Microbiology and Parasitology, University of Buea, Buea, South West region, CAMEROON|

**CONTROL ID:** 3367726

**TITLE:** Identifying Cancer Area in an Image using Scale-Mixing Spectra: a Wavelet Approach

**ABSTRACT BODY:**

**Abstract Body:** This work aims to show the behavior of different tissues in the same material by means of the time-varying spectral slopes. The methodology adopted to evaluate the cancer cells was based on an image approach, in particular using the two-dimensional scale-mixing wavelet transform. Wavelet transforms lead to coefficients representing the nature of a given signal at different locations/resolutions. From the transformed images several descriptive summaries derived. These descriptors involve time-varying slopes in wavelet spectra which are signatures of the degree of image regularity and fractality useful in the tissue classification. Images of neoplastic tissues were obtained through an optical method using the interference pattern formed when a biological material is illuminated by a laser. The static appearance of a speckle pattern is expressed by an image with clear and dark grains distributed over all the illuminated material. The speckle pattern expresses the boiling effect with the grains changing their shape and level of lightness related to the level of movement of the scatters of the coherent light. The irregular behavior of complex structures is difficult or impossible to quantify by standard modeling techniques, but when observations are inspected at different scales, there is in fact a regular relationship between the behavior among the scales. The methodology was applied to 128 interferometric images of an anaplastic mammary carcinoma in a female canine and in images of a basosquamous carcinoma in a feline (skin), obtained in regular time intervals in a rate of 0.08s. To implement the 2D-non-decimated wavelet transform, we consider 4 sub-images (64 x 64) in each of the 128 images, two from the cancer area and two from the healthy area. Initially we considered sub-images from the feline (skin) because it has more homogeneous areas than the canine (breast) that presented a high degree of infection even in healthy areas. Healthy area descriptors were found to have lower values than cancer area descriptors resulting in higher Hurst exponents. We observed for canine tissues that even in areas considered healthy but with high infection (characterized by whitish areas), it was possible to verify similar results to that found for the feline. The ability to apply the methodology in the animal tissues can be considered a potential usage in this research field.

**AUTHORS/INSTITUTIONS:** T. Safadi, Estatística, Universidade Federal de Lavras, Lavras, Minas Gerais, BRAZIL|B. Vidakovic, Georgia Tech University, Atlanta, Georgia, UNITED STATES|

**CONTROL ID:** 3367729

**TITLE:** Personalised screening schedules for optimal prevention of cardiovascular disease

**ABSTRACT BODY:**

**Abstract Body:** Cardiovascular disease (CVD) population screening strategies aim to identify and treat people at high risk of CVD. Current UK guidelines for CVD risk assessment recommend screening adults over 40 years old every 5 years and prescribing statins for those with a predicted 10-year CVD risk greater than 10%. The main goal of this work is to improve upon current screening practices, by providing a personalised screening schedule for each person, considering their specific risk profile.

We develop a problem-specific utility function, accounting for: (i) event-free life years, (ii) cost to the health service of providing statins for that individual and cost of screenings, (iii) undetected time (time that a patient is eligible for statins under the guidelines but he/she is actually not prescribed). The idea of measuring benefit and costs of statin prescription in terms of event-free life years was proposed in [1], but with a different goal: comparing prognostic models.

We optimize the previously defined utility function at different ages and according to 5-year CVD risk (low, intermediate and high risk strata). To assess 5-year CVD risk for each person and to properly adjust for time varying endogenous covariates, we use a two-stage dynamic landmark model [2]. The first stage consists in fitting at each landmark age (i.e., 40,45,...,80 years) a multivariate linear mixed effect model with random intercepts and slopes. Using this model, we are able to predict risk factor values (i.e., cholesterol, blood pressure and smoke) at different ages. The second stage consists in predicting the CVD risk through a Cox model, adjusted for the risk factor values estimated at stage one.

To conclude, in order to show the possible impact on CVD prevention strategy, we apply the proposed model to data from the Clinical Practice Research Datalink (CPRD), comprising primary care Electronic Health Records from the UK.

[1] Rapsomaniki, E., White, I. R., Wood, A. M., & Thompson, S. G. (2012). A framework for quantifying net benefits of alternative prognostic models. *Statistics in medicine*, 31(2), 114-130.

[2] Paige, E., Barrett, J., Stevens, D., Keogh, R. H., Sweeting, M. J., Nazareth, I., Petersen, I. & Wood, A. M. (2018). Landmark models for optimizing the use of repeated measurements of risk factors in electronic health records to predict future disease risk. *American journal of epidemiology*, 187(7), 1530-1538.

**AUTHORS/INSTITUTIONS:** F. Gasperoni, P. Newcombe, C. Jackson, J. Barrett, MRC-Biostatistics Unit, University of Cambridge, Cambridge, UNITED KINGDOM|A.M. Wood, University of Cambridge, Cambridge, UNITED KINGDOM|

**CONTROL ID:** 3367736

**TITLE:** A Genome-Wide Genetic Interaction Analysis for Lung Cancer Susceptibility using Machine-Learning Approaches

**ABSTRACT BODY:**

**Abstract Body:** Background: Although genome-wide association studies (GWAS) have studied genetic influences to complex diseases and have identified thousands of associations, still limited is the knowledge we possess of genetic interactions regarding disease susceptibility. Few GWAS have focused on pairwise interactions among SNPs that influence disease risks.

Methods: Machine learning applications can define how SNPs jointly influence disease risks through complex interactions. Tree-based machine-learning applications such as classification and regression trees (CART) and random forest (RF) methods are popular and convenient tools for understanding genetic interactions influencing disease development. Here we apply these methods to elucidate the higher-order interactions that influence lung cancer risk. We applied tree-based approaches using 18,444 lung cancer cases and 14,027 healthy controls from lung cancer OncoArray lung GWAS data. We first selected the SNPs very significantly associated ( $P < 1.0 \times 10^{-5}$ ) with lung cancer risk. RF, which consists of systematically fitting classification trees, was run 1,000 times to identify the most influential SNPs. Subsequently we applied CART to summarize and visualize interactions that predict risk.

Results: The final parsimonious tree included effects from genetic variants in CHRNA5, CLPTM1L, ZNRD1ASP, HCG9, TERT, CHRN4, and DNAJC5 for overall lung. In addition, we performed the tree-based interaction analyses for histological subtypes of lung carcinoma. The final tree showed the combination of genetic effects in or near ATM, CLPTM1L, TERT, FSTL5, and DCTN4 for lung adenocarcinoma, CHRNA5, MRPL21, HLA, CASP8, and TAP2 for lung squamous cell carcinoma, and finally, LHX5-AS1, CHRNA5, MICA, POT1, IRX4, KCNG1, MIR4790, LOC101928298, and CCHCR1 for lung small cell carcinoma.

Conclusion: Our results confirmed associations with CHRAN5, TERT, and HLA observed in previous study (McKay et al., 2017). Machine learning methods in genomics provide some benefits over logistic regression model with respect to identifying subgroups at higher risk of lung cancer development on the basis of genetic characteristics. Finally, with a greater appreciation of the complexities of genetic interaction, this study will converge upon the potential implications for clinical and translational research.

**AUTHORS/INSTITUTIONS:** J. Byun, Y. Han, C.I. Amos, Institute for clinical and translational research, Baylor College of Medicine, Houston, Texas, UNITED STATES|

**CONTROL ID:** 3367751

**TITLE:** Empirical Bayesian Disease Mapping of reported cases of malaria and diarrhea for children under age 5 using Poisson – Gamma and Poisson – Gamma – Rayleigh Models

**ABSTRACT BODY:**

**Abstract Body:** Reports by UNICEF indicate that under-five mortality rate for Nigeria is 100.2 per 1000 live births. Of these, malarial is responsible for about 25 per cent of the mortality. Diarrhea is also a major concern by health authorities regarding under-five mortality rate. To assess the intervention outcomes of various Governmental and Non-Governmental agencies, disease mapping is a viable statistical tool. The study applies the Poisson – Gamma and the Poisson – Gamma – Rayleigh models. Reported cases of malaria and diarrhea for children under age five are used in the study. Results show a mixed outcome at the regions while there has been consistent reduction in the average number of cases at the national level.

**AUTHORS/INSTITUTIONS:** I.A. ADELEKE, Actuarial Science & Insurance, University of Lagos, Lagos, NIGERIA|E.E. AKARAWAK, Mathematics, University of Lagos, Lagos, NIGERIA|A.U. MBATA, Mathematics, University of Lagos, Lagos, NIGERIA|

**CONTROL ID:** 3367771

**TITLE:** Nutrition label use is related to chronic conditions among Mexican adults

**ABSTRACT BODY:**

**Abstract Body:** Based on the Mexican National Health and Nutrition Survey of 2016 (ENSANUT MC 2016), the prevalence of overweight and obesity among the Mexican adult population ( $\geq 20$  years) is 72.5%. Likewise, the prevalence of diabetes and hypertension in this population is 9.4% and 25.5% respectively. These chronic, non-communicable diseases are related to poor diet habits. Processed and ultra-processed foods, which are high in calories and low in beneficial nutrients, are increasingly being eaten in Latin America. These products mandatorily display a nutrition label to help people make healthier food choices. The aim of our study was to investigate if people with chronic conditions (overweight, obesity, diabetes, and hypertension) make use of nutrition labels. We studied 5,013 Mexican adults aged 20 to 70 years interviewed in the ENSANUT MC 2016. We assumed that the participants used the nutrition labels if they answered affirmatively the question: "Do you read the nutrition label on packaged food and beverages when you shop?". A logistic regression model was used to examine the association between the use of nutrition labels and the type and number of chronic conditions reported by the survey participants while controlling for the effect of confounding variables. Only 40.9% participants used the nutrition labeling. We found that healthier participants were more likely to use nutrition labels compared to adults with a chronic condition. This result is contrary to what has been reported before in similar studies. In the presentation we will try to explain the reasons for this result and the limitations of our study.

**AUTHORS/INSTITUTIONS:** R. Aguirre-Hernandez, Facultad de Medicina, Departamento de Farmacologia, Universidad Nacional Autonoma de Mexico, Mexico City, MEXICO|C. Nieto Orozco, L. Tolentino-Mayo, E. Monterrubio-Flores, C. Medina, S. Barquera, Mexican National Institute of Public Health, Mexico City, MEXICO|S. Rincon-Gallardo, Human Nutrition, Foods, and Exercise, Virginia Tech, Virginia, Virginia, UNITED STATES|

**CONTROL ID:** 3367775

**TITLE:** Integrative analysis for identifying regulatory genetic variants in cancer patients

**ABSTRACT BODY:**

**Abstract Body:** In cancer development, it is crucial that non-coding genetic variants regulate gene expression, via altering transcription factor (TF) binding to DNA sequence. Predicting TF binding changes due to genetic variants can help identify genetic mutations and TFs potentially responsible for cancer. We develop efficient and scalable statistical test to evaluate regulatory power of genetic variants, based on an importance sampling technique coupled with a Markov model for background sequence generation. It quantifies TF motif matches to both reference and variant alleles and assess variant-led changes in TF motif matches. Small insertions or deletions of bases (InDels) as well as single-nucleotide polymorphisms (SNPs) can be tested with the method. Next, we conduct integrative analysis of genetic variants in leukemia patients and create a set of candidate genetic variants causal to leukemia with corresponding influential TFs. The integrative approach effectively filters disease-influential TFs and genetic mutations by combining the testing results with variant prioritization scores and TF binding profiles from sequencing experiments.

**AUTHORS/INSTITUTIONS:** S. Shin, Q. Zhou, University of Texas at Dallas, Richardson, Texas, UNITED STATES|Y. Zhang, J. Xu, UT Southwestern Medical Center, Dallas, Texas, UNITED STATES|

**CONTROL ID:** 3367809

**TITLE:** Iterative Least-squares Regression with Censored Data: A Survival Ensemble of Learning Machine

**ABSTRACT BODY:**

**Abstract Body:** Dealing with modeling for high-dimensional censored data is challenging because of the complexities in data structure. Many variable selection methods have been proposed for high-dimensional survival data for accelerated failure time (AFT) model. The study attempts to focus on extending variable selection procedure for censored high-dimensional data with AFT models using survival ensemble of popular machine learning techniques. Particularly, we modified the iterative least squares estimation technique as proposed by Jin et al. (2006) for AFT models by a survival ensemble of random forest and boosting machine learning techniques for obtaining precise estimation and variable selection. The implementation of these machine learning tools has been developed in light with a recent work by Khan and Shaw (2016). The performance of proposed methods has been demonstrated with high-dimensional censored data through a number of simulation examples and with a microarray data known as Diffuse Large-B-cell Lymphoma (DLBCL) where selection of genes that are linked with the survival time of DLBCL patients are studied. The proposed methods were compared with two similar methods in literature known as the modified resampling-based Buckley–James method (MRBJ) and Buckley–James Dantzig selector (BJDS) both developed by Khan and Shaw (2016). The simulation studies demonstrate very satisfactory variable selection performances for the proposed methods. The proposed boosting and random forest based methods outperform existing methods for most of the cases. The DLBCL data analysis also suggests that both proposed methods are able to find the important genes that are related to survival of patients and also can predict the survival time of future patients with small prediction error. The proposed methods are easy to understand and they perform estimation and variable selection simultaneously.

**AUTHORS/INSTITUTIONS:** M. Khan, N. Hossain, Institute of Statistical Research and Training, Applied Statistics, University of Dhaka, Dhaka, Dhaka, BANGLADESH|

**CONTROL ID:** 3367821

**TITLE:** Choice between Mixed and Multiplicative

Models in Time Series

Decomposition.

**ABSTRACT BODY:**

**Abstract Body:** This work discusses the condition under which the mixed model best describes the pattern in an observed time series data, while comparing it with those of the additive and multiplicative models. Existing studies have focused on how to choose between additive and multiplicative models, with little or no emphasis on the mixed model. The ultimate objective of this study is therefore, to propose a statistical test for choosing between mixed and multiplicative models when the trending curve is linear in descriptive time series analysis. The method adopted in this study is the Buys-Ballot procedure developed for choice of model by [1]. Results show that the column/seasonal variance of the Buys-Ballot table is, for the mixed model, a constant multiple of the square of seasonal effect and for the multiplicative model, a quadratic (in  $j$ ) function of the square of the seasonal effects. Therefore, test for the choice between mixed and multiplicative models has been proposed based on the column/seasonal variances of the Buys-Ballot table have been used to illustrate the applicability of the proposed test. Using empirical examples, the proposed test statistic identified the mixed model correctly in 98 out of the 100 simulations.

**AUTHORS/INSTITUTIONS:** H.L. Mbachu, Statistics, Imo State University Owerri, Imo State Nigeria, Owerri, Imo State, NIGERIA]

**CONTROL ID:** 3367831

**TITLE:** Water quality monitoring of streams in Puerto Rico using GAMM

**ABSTRACT BODY:**

**Abstract Body:** Water quality monitoring is a crucial process to assess the health of an ecosystem. Decisions on natural resource management must be based on current and historical conditions, and the effect of policy changes on the water quality must be carefully assessed comparing water quality and trends before and after the changes. In this paper we analyze 57 streams in Puerto Rico to study the trends over time of water quality parameters such as pH, water temperature, dissolved oxygen concentration, phosphorus concentration, etc. Data were obtained from US Geological Service records and several additional studies conducted in Puerto Rico in the last 15 years. The data available span from 1958 through 2019, although there are significant gaps in several years. Generalized additive mixed models were able to capture the most important trends using splines while taking into account the longitudinal nature of these data using appropriate random effects. It also incorporated meaningful covariates to study differences between regions, impact by water treatment plants, and the effect of elevation and other geographical features. This modeling environment also provided an appropriate framework to test the significance of trend changes at important times, and to predict current trends in water quality parameters.

**AUTHORS/INSTITUTIONS:** R.E. Macchiavelli, M. Vazquez, G. Martinez, Agroenvironmental Sciences, University of Puerto Rico, Mayaguez, PUERTO RICO|C. Perdomo, Mathematical Sciences, University of Puerto Rico , Mayaguez, PUERTO RICO|

**CONTROL ID:** 3367843

**TITLE:** Extending SIMEX for Non-zero Mean Covariate-Dependent Measurement Error to Parametrically Account for Multiple Continuous Covariates

**ABSTRACT BODY:**

**Abstract Body:** Administrative databases present great advantages for research on drug safety and effectiveness. Yet, these data were not collected for research purposes which may thus lead to many statistical challenges. In particular, to determine individuals' exposures to drugs, researchers rely on drug prescriptions or dispensations from pharmacies. However, due to non-adherence, these may not accurately represent the actual drug intake. Such discrepancies induce measurement error (ME) in the assessment of the true drug exposure, which is known to result in biased inference in naive analyses.

In the absence of validation data, a simple approach to reduce bias due to ME is the simulation-extrapolation method (SIMEX). For classical additive ME, SIMEX only requires distributional assumptions about the variance of the ME, as it assumes the errors have zero mean. However, there are many instances in which errors may not have zero mean or may depend on other observed covariates. Recent methodological developments have proposed SIMEX extensions to account for non-zero mean covariate-dependent ME. However, applicability of these approaches becomes quickly limited to the case in which the error distribution depends solely on categorical variables. It also requires the analyst to correctly supply the mean of the error distribution in every subset of the population defined by the covariates, thus relying on in-depth knowledge of the error distribution, which may not be realistic.

We propose to extend the above method to allow to model parametrically the mean of the unknown error distribution. This improves upon existing methods by accounting for dependencies of the ME on many, possibly continuous, variables. It also does not require any additional knowledge of the ME distribution.

We use a simulation study in which we mimic various real-life non-adherence patterns to evaluate the performance of the method to correct for this type of ME. We also assess the sensitivity of the method to various tuning parameters in the SIMEX procedure and compare it to a standard regression calibration approach. Preliminary results show that naive methods produce biased estimates. The proposed method performs well in simulation, by reducing the bias due to ME. Finally, we demonstrate the performance of the suggested SIMEX when the sample size is relatively small.

**AUTHORS/INSTITUTIONS:** S. Ferreira Guerra, M. Abrahamowicz, R. Platt, McGill University, Montreal, Quebec, CANADA|

**CONTROL ID:** 3367849

**TITLE:** An ensemble method for RNA-Seq differential analysis

**ABSTRACT BODY:**

**Abstract Body:** RNA sequencing is now widely used in biological and biomedical research for identifying important genes associated with diseases. Currently, there are many open source RNA-Seq differential analysis software available. However, when the sample size of the RNA-Seq experiment is small, the false discovery rate (FDR) is not well controlled for all the current methods. The inflated FDR makes further validation of selected significant genes prone to high probability of false discoveries and waste the time and resources used for experimental validations. To reduce the FDR in RNA-Seq differential analysis with small sample sizes for many biomedical researchers struggling with funding and resource limitations, we proposed an ensemble method to combine current RNA-Seq differential analysis methods through the proposed novel weighting algorithm, and obtain a p-value for each gene using the marginal distribution of the final weighted rank scores generated using permutations to rank all genes. We conducted simulation studies to show the significant decrease in estimated FDR using our proposed ensemble method. Real RNA-Seq data examples were also used to compare the performance of our proposed ensemble method with existing popular RNA-Seq differential analysis methods.

**AUTHORS/INSTITUTIONS:** D. Li, Z. Xie, Clinical and Translational Research, University of Rochester Medical Center, Rochester, New York, UNITED STATES|G. Yu, University at Buffalo, Buffalo, New York, UNITED STATES|A. Paine, University of Rochester Medical Center, Rochester, New York, UNITED STATES|

**CONTROL ID:** 3367875

**TITLE:** Robust Estimation of the Survival Curve at Interim

**ABSTRACT BODY:**

**Abstract Body:** Adaptive designs provide an attractive possibility of changing study design parameters, such as the trial's sample size, in an ongoing trial. There are still many open questions with respect to adaptive designs for time-to-event data. Among other aspects, this is because survival data, unlike continuous or binary data, is not directly observable after the end of a treatment.

Evaluating survival data at interim analyses usually includes a patient overrun since the recruitment is commonly not stopped at the interim time point for logistic reasons. Moreover, the timing of the interim analysis is a crucial point to save patients on the one hand, which is only possible during the recruitment phase, but to also build decisions upon a reasonable level of information, which requires a certain amount of follow-up observation.

To recalculate the required sample size based on interim time-to-event data, an updated hazard ratio estimation along with an estimator of the estimated proportions of events at a specific time point is required. Therefore, a parametrization of the full underlying event-time distribution is needed. This can possibly be estimated from the interim data if the form of the class of the parametric distribution is known. This however, is usually not the case and moreover the number of observed events at interim is limited and the survival curve at interim is truncated by the interim time point.

In this work, we investigate approaches for a robust estimation of the survival curve at interim with as few parametric assumptions as possible. This will be done by comparing several parametric survival model estimators and their ability to fit the actual data in an automated way.

**AUTHORS/INSTITUTIONS:** N. Akbari, Charité - Institut für Biometrie und Klinische Epidemiologie, Berlin, GERMANY|

**CONTROL ID:** 3367879

**TITLE:** A graded response model for mental health status in Sao Paulo state, Brazil

**ABSTRACT BODY:**

**Abstract Body:** Mental disorders, in particular depression, have become a highly prevalent disease in contemporary society, according to the World Health Organization. The number of cases of depression increased by 18% between 2005 and 2015. In Brazil, the prevalence estimate is 5.8% of the population. The diagnosis is difficult because of the lack of an objective method and the specialist relies on the analysis of historical events and application of mental health questionnaires devised to assess patient depression status. Thus depression status fits well within the latent variable modeling framework and the item response theory, which considers the modeling of continuous variables, is being used to explore and identify the status of the diseases.

The Brazilian National Health Survey (a partnership between IBGE and the Ministry of Health) evaluated the health and lifestyle of the adult population in 2013. Perception of mental health is included through the application of the Patient Health Questionnaire, the PHQ-9. It consists of nine questions with Likert-scale answers, in which each participant reports the frequency of symptoms she/he experienced in the 15 days before the survey, with answers in four categories: never; less than of the half days; more than half of the days and almost every day.

We applied the gradual response model (GRM) to explore the disease status in the state of São Paulo (n=5305). Although the survey used a nationwide complex sampling system, in this preliminary work we restricted to data from São Paulo state and did not consider the actual design.

The fitting of the GRM allowed the construction of a scale measuring the severity of depression, the latent trait. The scale is arbitrary but the ordering of the values is important, higher values indicating more severe condition. For our data, the scale ranges from -3 to 6 with negative values associated with individuals that reported almost no symptoms. The more informative items for the range from 0.5 to 2.75 are "disinterest in doing things", "depressed mood" and "feeling bad about yourself". "Suicidal ideation" contributes to the discrimination of individuals in the interval from 1.75 to 3.75. At this stage, this analysis is exploratory. Improvements are sought towards combining response categories and incorporation of the sampling design.

**AUTHORS/INSTITUTIONS:** L. Ragoatto, L.A. Trinca, Bioestatística, Unesp, Botucatu, Sao Paulo, BRAZIL|

**CONTROL ID:** 3367883

**TITLE:** Why Students Hate Biometry and Why it Matters to the Lectures

**ABSTRACT BODY:**

**Abstract Body:** Biometry/Statistics is a subject that many students have to take, and most of them find uninteresting, unpopular and feedback from students indicates that some find it more difficult than other subjects. This student frustration is generally offloaded onto the lecturers and since student evaluations are considered as “predictors of learning” finally it ends up with lower evaluation scores of lectures. Furthermore, it is interesting that within the same class individual student comments on the lecturer are varied wildly and have negative to positive attitudes about their interest in statistics and the majority of students’ feelings towards statistics are less positive. This negative attitude is increasing but statistics is evolving and growing. New resources for teaching statistics and good teaching practices increase learning experience, encourage students and ensure a more positive experience. The overall experience and evaluations show that student attitude towards statistics has changed but not entirely satisfactory. One important thing we need to address is the association between students’ past experiences in mathematics as well as their attitudes towards statistics to reduce their feelings of fear and frustration. Furthermore, it is important to build a platform that brings teaching people closer together to share their experience of teaching various statistical concepts and methods applied to promote and increase the student experience and discuss how to change the students’ experience.

**AUTHORS/INSTITUTIONS:** B. Dayananda, School of Agriculture and Food Sciences, University of Queensland, Brisbane, Queensland, AUSTRALIA|

**CONTROL ID:** 3367890

**TITLE:** Statistical Methods for Addressing Challenges in Medication Dose Extraction from Electronic Health Records

**ABSTRACT BODY:**

**Abstract Body:** For large-scale population studies concerning the associations between drug exposure and clinical outcomes/phenotypes, electronic health records (EHR) can be rich source of medication dose data. However, obtaining medication dosing information from EHRs involves multiple challenges. Firstly, as structured dose data obtained from EHRs are often incomplete, medication dosing information should be abstracted from clinical notes using a natural language processing (NLP) system. However, existing NLP systems may not provide sufficiently accurate dose data for medication-based studies. Secondly, even if medication entities could be accurately extracted from clinical notes using an NLP, relevant entities should be paired up in order for them to serve as medication dose data in further data analysis. This process is challenging and requires a sophisticated post-processing algorithm. Thirdly, often extracted medication dose data are not unique for each day, yielding conflicting multiple dose information. We will discuss these challenges and statistical methods that can help to address these challenges in medication research using EHRs.

**AUTHORS/INSTITUTIONS:** L. Choi, C.L. Beck, H.L. Weeks, E. McNeer, Biostatistics, Vanderbilt University, Nashville, Tennessee, UNITED STATES|M.L. Williams, C.A. Bejan, J.C. Denny, Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, UNITED STATES|

**CONTROL ID:** 3367893

**TITLE:** A class of two-sample binary and survival statistics with application to immunotherapy trials

**ABSTRACT BODY:**

**Abstract Body:** Cancer immunotherapies have emerged as alternative treatments to fight cancer. The response pattern in Immuno-Oncology (IO) therapy differs from previous cancer treatment regimes and is mainly characterized by a delay in the clinical effect<sup>1</sup>. As a consequence, traditional endpoints such as objective response --based on tumor shrinkage-- and overall survival might not capture the benefits of IO. Moreover, the assumption of proportional hazards seldom holds because of the delayed separation of the survival curves. Hence, novel endpoints and alternative statistical approaches accounting for the non-proportionality of the hazards are needed.

Aiming to capture both tumor and survival responses of the immuno-response, we propose a two-arm trial design where the efficacy is evaluated using a short-term binary endpoint and a survival endpoint. This design would provide an insight of the tumor activity through the binary endpoint while the study continues with the time-to-event response.

We present a novel class of statistics for testing the equality of proportions and the equality of survival functions. We build our proposal on a weighted combination of a score test for the difference in proportions and a Weighted Kaplan-Meier statistic-based<sup>2</sup> for the difference of survival functions.

The proposed statistics are fully non-parametric and do not need the proportional hazards assumption for the survival outcome. We present the asymptotic distribution of these statistics, propose a variance estimator and outline their asymptotic properties. We discuss different choices of weights including those that control the relative relevance of each outcome and emphasize the type of difference to be detected in the survival outcome. We evaluate our proposal in terms of the significance level with a simulation study. We illustrate our proposal through an IO trial dataset. All the required functions to use these statistics have been integrated into an R package.

1. Anagnostou V, Yarchoan M, Hansen AR, et al. (2017). Immuno-oncology trial endpoints: Capturing clinically meaningful activity. *Clinical Cancer Research*. 23(17):4959-4969.

2. Pepe MS, Fleming TR (1989). Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data. *Biometrics*. 45(2):497-507.

**AUTHORS/INSTITUTIONS:** M. Bofill Roig, G. Gomez Melis, Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, SPAIN|

**CONTROL ID:** 3367916

**TITLE:** Integrated Analysis of Microarray Data to Investigate Specific Gene Modules under Different Hormone Treatments by subCID

**ABSTRACT BODY:**

**Abstract Body:** Abiotic and abiotic stresses impact crop productivity. The scenario is expected to be severed due to the climate change. Thus, it rises the inquires of stress tolerant or resistant crop varieties. This has led to this study of understanding the tolerant or resistant mechanisms in gene-regulation level. Particularly, the gene-regulatory mechanisms of plant hormones were of interest because it is widely acknowledged that plant hormones play a critical role in the metabolism and physiological regulation of plant cells when plants are under stresses. We had first performed the weighted gene co-regulatory network analysis (WGCNA) to identify the gene modules that are possibly co-regulated under different hormone treatment. The multivariate subCID values under different hormone treatments for the gene modules were utilized to further investigate the specificity of the gene modules. The biplots based on the singular value decomposition on the subCID values helped the molecular biologists to design finer experiments to validate our findings. The results of the illustrative data in this study may elucidate stress-specific mechanisms in plants, which have the potential to accelerate the development of crop production strategies for coping with climate change in the future.

**AUTHORS/INSTITUTIONS:** L.D. Liu, Y.N. So, Department of Agronomy, National Taiwan University, Taipei, TAIWAN|

**CONTROL ID:** 3367919

**TITLE:** A practical guideline for designing a phase I oncology clinical trial

**ABSTRACT BODY:**

**Abstract Body:** The primary object of a phase I oncology trial is to determine a recommended dose of the study agent for phase II trials. This recommended dose is the maximum tolerated dose (MTD) for cytotoxic agents. Several novel designs for phase I oncology trials have been introduced in recent years; Model-assisted designs such as mTPI, BOIN, and keyboard design have recently been proposed and gained popularity. These designs allow generation of dose escalation/de-escalation rules similar to the 3+3 designs prior to the trial, thus intuitive and easy to implement with user-friendly software. These novel designs have comparable general performance. Therefore, it is often not clear which one best fits the needs of a specific phase I oncology clinical trials. Additionally, there is not a good answer to the sample size, cohort size and maximum number of subjects per dose level. In this paper, we will take a closer look at each of these novel designs and use extensive simulation studies to provide answers to the key questions: "Why this design?" and "How many patients do I need?" The proposed practical guidelines should increase the adoption of these novel designs for phase I oncology clinical trials.

**AUTHORS/INSTITUTIONS:** J. Lim, Y. Chen, A. Collins, C. Bradbury, Knight Cancer Institute, Oregon Health and Science University, Portland, Oregon, UNITED STATES|

**CONTROL ID:** 3367921

**TITLE:** Evaluating Random Survival Forest: Small Sample Operating Characteristics

**ABSTRACT BODY:**

**Abstract Body:** In an era of new computing technologies, many learning algorithms have been developed. Most of those algorithms are for bigger and complex varieties of available data. Medical researchers are fascinated by ensemble learning that can be extended to the models including right censored survival. Random survival forest (RSF) was popularly adopted among many clinical and basic researchers without knowing small sample operating characteristics as open source software are available at no cost.

We conducted a retrospective analysis of patients with chronic lymphocytic leukemia (CLL) treated with either chemo-immunotherapy (CIT) or kinase inhibitors at 10 US academic medical centers between 2000-2018. Outcomes in CLL are highly variable and influenced by both biologic and clinical factors. The Cumulative Illness Rating Scale (CIRS) is frequently used to assess comorbidities in CLL. Our group has demonstrated that CIRS correlates with survival in patients treated with either CIT or ibrutinib. Yet, CIRS has not become part of common clinical practice due to complexities in scoring since 14 comorbidities need to be evaluated. Investigators are interest in reducing dimensionality of CIRS to enhance practicality and encourage adoption in clinic. They also want to confirm reliability of the resulted predictive model through internal validation.

Event free survival was model with supervised random survival forests (RSF) for identifying most importance comorbidities in the presence of known prognostic factors. Patients were randomly divided into a training-set (n=381) and validation-set (n=189). To assess the robustness of RSF for this relatively small data set, we randomly generated 100 different training sets. For each training set, we fit 50 RSFs, each comprised of 1000 survival trees. We examined the distribution of variable importance and minimal depth to select the best subset of comorbidities. We found that this chosen subset was greatly varied by training-sets. We obtained the same subset only 47% of the time, thus RSF variable selection approach showed consistency less than half of the time. This motivates us to make a close investigation of small sample properties of RSF. Simulation studies were performed to access operating characteristics of RSF, to find the minimal sample size to achieve a specified consistency threshold.

**AUTHORS/INSTITUTIONS:** B.S. Park, A. Kaempf, M. Gordon, A. Danilov, Knight Cancer Institute, Oregon Health and Science University, Portland, Oregon, UNITED STATES|B.S. Park, School of Public Health, Oregon Health & Science University, Portland, Oregon, UNITED STATES|

**CONTROL ID:** 3367926

**TITLE:** Individual Trends in the Loss of Bone Mineral Density

**ABSTRACT BODY:**

**Abstract Body:** The rate of bone mineral density (BMD) loss affects the length and quality of lives of a large proportion of elderly people. The BMD loss contributes to the development of osteoporosis in post-menopausal women and aged men and is a strong risk factor for bones fractures and its related mortality. In the following work, our primary objective is to investigate the association between genetic information of patients and their tendency to the loss of BMD, i.e. their vulnerability to the osteoporosis. To address this we examine different characterisations of the components which influence the loss of BMD, discerning between life-cycle accumulated exposures, individual determinants as well as cohort effects. We assume that the loss of BMD varies stochastically around its population curve overage. By the addition of individual and cohort components, we investigate the stochastic long-term trend and level of BMD. The cohort effect is used to capture implications brought by environmental, social and economic circumstances inclusive to a sub-population born in a particular calendar year. It is shown that various model assumptions impact on predictors of an individual tendency to the loss of BMD, and, consequently, provide different associations with available genetic information. The findings are illustrated on data available from the longitudinal Dubbo Osteoporosis Epidemiology Study.

**AUTHORS/INSTITUTIONS:** T. Nguyen, Genetic Epidemiology of Osteoporosis Lab, Garvan Institute of Medical Research, Sydney, New South Wales, AUSTRALIA|L. Ryan, M.P. Wand, Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers, Parkville, Victoria, AUSTRALIA|T. Nguyen, Clinical School, St Vincent's Hospital, Faculty of Medicine, The University of New South Wales, Sydney, New South Wales, AUSTRALIA|T. Nguyen, School of Medicine Sydney, University of Notre Dame Australia, Sydney, New South Wales, AUSTRALIA|T. Nguyen, School of Biomedical Engineering, University Technology Sydney, Sydney, New South Wales, AUSTRALIA|D. Toczydlowska, L. Ryan, M.P. Wand, School of Mathematical and Physical Sciences, Faculty of Science, University Technology Sydney, Sydney, New South Wales, AUSTRALIA|

**CONTROL ID:** 3367932

**TITLE:** A Bayesian latent class approach to causal inference with longitudinal data

**ABSTRACT BODY:**

**Abstract Body:** Causal inference from longitudinal data with a large number of time-dependent covariates are challenging, especially from a Bayesian perspective. Bayesian causal methods that follow a parametric specification of the treatment assignment model, the causal outcome model or the joint likelihood of treatment, outcome and covariates, are analytically intractable and often unappealing when faced with high-dimensional confounders. One possible approach for dimensionality reduction is to model the set of confounders as class indicators in a latent class analysis - interpreting the latent class membership as a confounder in the causal framework. This approach mimics the treatment assignment process often seen in observational studies with administrative data that contain a large number of variables which are indicative of the patient's disease and health status. Treatment assignment under this design follows a clinician-driven decision process, where the treating clinician determines the treatment option based on their classification of the patient's health status using these high-dimensional indicators. In this paper, we consider a causal effect that is confounded by an unobserved, visit specific, latent class in a longitudinal setting. This latent class can be viewed as a time-dependent disease-risk/comorbidity class that can be identified through a set of observed indicators. We formulate the joint likelihood of the treatment, outcome and latent class models conditionally on the class indicators, which permits a full Bayesian estimation of the causal effects. A simulation study is conducted to examine the performance of our proposed method with different levels of confounding and varying numbers of class indicators, and our approach is illustrated through a study of the effectiveness of intravenous immunoglobulin therapy in treating newly diagnosed juvenile dermatomyositis.

**AUTHORS/INSTITUTIONS:** K. Liu, O. Saarela, G. Tomlinson, E. Pullenayegum, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, CANADA|K. Liu, E. Pullenayegum, Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, Ontario, CANADA|G. Tomlinson, Medicine, University Health Network, Toronto, Ontario, CANADA|

**CONTROL ID:** 3367933

**TITLE:** Multiclass machine learning classification of functional brain images

**ABSTRACT BODY:**

**Abstract Body:** In this study, we analyze a dataset containing functional brain imaging from 6 normal healthy controls, and 196 Parkinson's disease (PD) patients that are divided into 5 stages according to the severity of illness. The goal is to predict patients' PD illness stages via their functional brain images. Used approaches include multivariate statistical methods (linear discriminant analysis (LDA), support vector machine (SVM), decision tree (DT), multi-layer perceptron (MLP)), ensemble learning models (random forest (RF), adaptive boosting (AdaBoost)), and deep convolutional neural network (CNN). For statistical and ensemble models, various dimensional reduction approaches (principal component analysis (PCA), multilinear principal component analysis (MPCA), intensity summary statistics (IStat), Laws' texture energy measure (LTEM)) are performed to extract features, synthetic minority over-sampling technique (SMOTE) is used to deal with imbalanced data problem, and the best combination of hyper-parameters is found by grid search. For CNN modeling, we apply image augmentation technique to increase and balance data sizes over different disease stages, we adopt transfer learning idea to bring pre-trained VGG16 weights and architecture into model fitting, and we also try out a state-of-the-art machine learning model that can generate an optimal neural architecture automatically (AutoML). It is found that LDA and RF are the analytic approaches with the highest prediction accuracy rate for traditional and ensemble models, respectively. Overall, the deep CNN model outperforms other approaches, which can capture significant features from imaging, have a better separation of normal controls and patients with PD, and reach higher classification accuracy.

**AUTHORS/INSTITUTIONS:** G. Huang, C. Lin, Y. Cai, Institute of Statistics, National Chiao Tung University, Hsinchu, TAIWAN|

**CONTROL ID:** 3367937

**TITLE:** Bayesian Reversible Jump MCMC for Spatially constrained Skew Normal Mixture Model in Application of MRI Brain Tumor Segmentation

**ABSTRACT BODY:**

**Abstract Body:** Brain tumor image segmentation is always challenging due to its unclear shape and boundaries. This condition leads to different data patterns in a clustering model. This study proposed an improved algorithm for model-based clustering through a finite mixture model by employing an adaptive distribution called Skew Normal Mixture Model (Skenomimo). The main contribution is a mixture model that could handle the different data patterns both symmetrical and asymmetrical. The proposed model also solved the restriction of the Gaussian Mixture Model (GMM), which overcomes the short tail characteristic problem to give a more parsimony model. Moreover, the spatial dependencies also applied to the proposed model for improving the robustness against noise. Model optimization is done by building the Bayesian Reversible Jump MCMC algorithm to provide the automatization for the optimum number of clusters. The model performance is evaluated against the GMM and other previous states of the art using the Correct Classification Ratio (CCR). Finally, this study achieved that the proposed model provides better segmentation results for MRI-brain tumor data set with the average of CCR is more than 97%.

**AUTHORS/INSTITUTIONS:** A.A. Pravitasari, Statistics, Universitas Padjadjaran, Bandung, West Java, INDONESIA| A.A. Pravitasari, N. Iriawan, I. Irhamah, K. Fithriasari, S.W. Purnami, Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, East Java, INDONESIA| W. Ferriastuti, Radiology, Universitas Airlangga, Surabaya, East Java, INDONESIA|

**CONTROL ID:** 3367939

**TITLE:** Evaluating the temporal trend in Gaussian spatio-temporal linear models

**ABSTRACT BODY:**

**Abstract Body:** Spatio-temporal models are well known in the literature, however to present statistical methods to evaluate the temporal dependence between observations in this type of model is still a challenging task. We propose covariance matrix perturbation scheme under local influence approach to check the assumption of temporal structure in Gaussian spatio-temporal linear models with repeated measures. We evaluate the presence of temporal structure in a soybean productivity data set collected at the same sites over five years, and so on the explanatory variables, the soil chemical contents. This methodology can be considered as a goodness of fit, verifying assumptions such as independence across time and heteroskedasticity. The results showed that there is no misleading in considering a model with independent repeated measures for this data set. Extension of this methodology to other distributions such as the elliptical family of distributions and to explore assumptions as well as separable covariance functions are straightforward.

**AUTHORS/INSTITUTIONS:** M.A. Uribe Opazo, Ciências Exatas e Tecnológicas, Universidade Estadual do Oeste do Paraná, Cascavel, Paraná, BRAZIL|F. De Bastiani, Statistics, Universidade Federal de Pernambuco, Recife, Pernambuco, BRAZIL|M. Galea, Pontificia Universidad Catolica de Chile, Santiago, CHILE|

**CONTROL ID:** 3367964

**TITLE:** A new data stream clustering algorithm for knowledge-driven clinical and physiological decision support systems

**ABSTRACT BODY:**

**Abstract Body:** Many medical dataset has the nature of data stream (online data) and has to find right information at the right time. For example, checking the cessation of breathing, it requires extensive monitoring to determine diagnosis. It has difficulties to visually assess the etiology and type of apnoea. Thus, continuous feature extraction based on data steam clustering can be the possible effective mean for the diagnosis, what we define as the knowledge discovery for the point of care applications.

Data stream clustering plays an important role in clinical support systems for knowledge extraction. It has increasingly become a ubiquitous and essential tool in medical data stream analysis for finding decision patterns. Many studies on the clustering algorithm has been developed for static datasets, however, they cannot be applied on data streams due to the uncertainty in volume, arrival speed, and time. Therefore, because of these uncertainties, a good data stream clustering algorithm needs to develop in order to achieve minimum processing delay, detection of noise, and getting less amount of memory.

In this article, experimental result shows the higher performance of the proposed data stream clustering algorithm over existing algorithms (CODAS, CEDAS, etc.) in terms of several performance criteria, such as cluster quality, noise sensitivity, processing speed, and dimensionality. Furthermore, We apply the proposed algorithm to real-world data stream, clinical and physiological data stream, to demonstrate its capability of handling content changes based on drift analysis in different terms (i.e., short-term, medium-term , and long-term ).

**AUTHORS/INSTITUTIONS:** M.M. Ahmed, Computer Science and Engineering, University of Barishal, Bangladesh, Barishal, Barishal Sadar, BANGLADESH|

**CONTROL ID:** 3367965

**TITLE:** Empirical simulation of very rare variant genetic data

**ABSTRACT BODY:**

**Abstract Body:** Simulating realistic rare variant genetic data is vital for accurate evaluation of new statistical methods. Research suggests that large sample sizes and functional information are necessary for sufficiently powered rare variant association tests. Further, the distribution of simulated variants should be similar to that observed in sequencing data. HAPGEN2 (Su 2011) accurately simulates common genetic variants, but is unable to simulate data that reflects the observed allele frequency spectrum (AFS) for very rare variants, such as singletons and doubletons. Currently there is no simulation software that produces large sample sizes with realistic functional annotation and the expected AFS across all, including very rare, variants.

We developed RAREsim, a flexible software that simulates large sample sizes of genetic data with an AFS similar to that observed in sequencing data. Because RAREsim simulates from a sample of real haplotypes, existing functional and other genetic annotation can be used, capturing known and unknown complexities of real data. RAREsim is a two-step algorithm. First, RAREsim simulates haplotypes using HAPGEN2 (Su 2011) allowing for mutations to occur at most sites across the region. Second, RAREsim prunes the rare variants using the expected number of variants at each minor allele count. The expected number of variants is calculated from an estimate of the total number of variants in the region and the AFS. Since the AFS and total number of variants have been shown to vary by ancestry and variant type (e.g. synonymous, intron), we provide tuning parameters to enable user flexibility while maintaining the general relationship between the number of variants, AFS, and sample size. While we derive default parameters from the Genome Aggregation Database (gnomAD), the user has the ability to vary the number and distribution of variants to reflect their desired distribution for the region.

RAREsim is available as an R package, with a Shiny App for the user to select and visualize the allele distribution parameters. RAREsim provides the ability to simulate large samples of rare variant data with functional annotation and the expected AFS for all, including very rare, variants. Realistic rare variant simulations are critical for rare variant method development. In turn, advances in these methods will allow for a greater understanding of the role rare variants play within health and disease.

**AUTHORS/INSTITUTIONS:** M.M. Null, A. Hendricks, Mathematical and Statistical Sciences, University of Colorado Denver, Denver, Colorado, UNITED STATES|A. Hendricks, Human Medical Genetics and Genomics Program, University of Colorado Anschutz Medical Campus, Aurora, Colorado, UNITED STATES|A. Hendricks, Biostatistics and Informatics, Colorado School of Public Health, Aurora, Colorado, UNITED STATES|J. Dupuis, Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, UNITED STATES|

**CONTROL ID:** 3367969

**TITLE:** Differential Networks Analysis via Scalable Bayesian Networks for Single Cell RNA-seq Data

**ABSTRACT BODY:**

**Abstract Body:** Single cell RNA sequencing is an emerging technology which has revolutionized a wide range of biomedical research. The technology, however, generates highly sparse data matrices with excessive zeros, posing statistical and computational challenges to analysis of differential gene regulatory networks. As for differential networks, only few methods are developed to examine single cell RNA sequencing (scRNA-seq) data, while many approaches have been proposed for bulk RNA sequencing data. We develop a novel framework for the analysis of differential gene regulatory networks by modeling asymmetric (directed) conditional dependence via Bayesian networks. We use Bayesian networks based on zero-inflated Negative Binomial models to deal with excessive zero observations in the data. Our use of zero-inflated Negative Binomial models allows our approach to account for not only zero-inflation but also overdispersion of scRNA-seq data. In order to detect differential patterns of gene regulations across experimental groups, we incorporate parameters into the model, which indicate differences in the strength of gene interactions across groups. Spike-and-slab priors are adopted to make direct inference on such differences as well as to induce sparsity in the edge selection. A Markov chain Monte Carlo algorithm is proposed for the posterior inference on parameters in our model. Simulation studies demonstrate that our method outperforms existing differential network methods for scRNA-seq data. We apply our method to a scRNA-seq dataset, which figure out significant and interesting differential gene regulation from the data.

**AUTHORS/INSTITUTIONS:** J. Choi, Y. Ni, Department of Statistics, Texas A&M University, College Station, Texas, UNITED STATES|

**CONTROL ID:** 3367976

**TITLE:** Parsing latent factors in high dimensional classification

**ABSTRACT BODY:**

**Abstract Body:** High throughput biological data often contains signals from multiple unobserved latent factors in addition to the signal of primary interest. In a classification analysis, some of these latent factors may be partially correlated with the phenotype of interest and therefore helpful, some may be uncorrelated and thus merely contribute additional noise, while more perniciously, some may be spuriously correlated with the phenotype in the training set but not in the target population, leading to poor generalized predictive performance. Moreover, whether potentially helpful or not, these latent factors may obscure weaker direct effects that capture the signal of primary interest. It is therefore desirable to separate out these latent variables. This talk has two parts. The first outlines a classification algorithm that first isolates the signal of primary interest from other latent factors, but then exploits both to improve prediction, leading to sometimes substantial gains. The second discusses how to remove uncorrelated, non-stationary, or otherwise harmful latent factors.

**AUTHORS/INSTITUTIONS:** J.A. Gagnon-Bartsch, Y. Pan, Statistics, University of Michigan, Ann Arbor, Michigan, UNITED STATES|

**CONTROL ID:** 3367977

**TITLE:** A Novel Bioinformatics Approach of Breast Cancer Diagnosis by MALDI-TOF Based Serum N-glycan Bio-signature

**ABSTRACT BODY:**

**Abstract Body:** Breast cancer is one of the most frequent causes of death for women worldwide. Since early detection of cancer enables efficient treatment and prevents loss of resources, numerous molecular diagnostic methodologies have been suggested as an alternative to previous methods such as mammography and ultrasonography. These range from exploiting protein biomarkers to analyzing nucleic acid sequences. However, molecule-based diagnosis strategies have frequently shown insufficient accuracy due mainly to data reproducibility. Here, we present a novel cancer diagnosis algorithm based on “library diagnostics” and “combination of categorization and bio-signature”. 19 N-glycans were identified from MALDI-TOF data as bio-signature for breast cancer, each one has its own computed importance (“weight”). Based on these N-glycan markers, we defined optimally classified, multiple distinguishable sub-groups from sample groups of both healthy controls and patients using an iterative algorithm. Furthermore, medical history factors showing significant correlations with sub-group classification were identified. Weight factor for each N-glycan marker was found to enhance diagnostics accuracy. As a result, our diagnosis model was able to distinguish healthy controls and breast cancer patients with an accuracy exceeding 80 %. With further optimization and adjustment, we expect our model to be applicable as a practical assistant tool for other cancer diagnostics as on breast cancer. In conclusion, categorization concept shall increase diagnostics accuracy, along with bio-markers.

**AUTHORS/INSTITUTIONS:** K. Lee, B. Seo, J. Lee, D. Kim, C. Ko, NosQuest Corp., Seongnam, KOREA (THE REPUBLIC OF)|E. Jo, ASTA Corp., Suwon, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3367987

**TITLE:** Current Situation and Consideration on Sample Size Re-estimation

**ABSTRACT BODY:**

**Abstract Body:** In order to develop medicines, treatments or medical devices, clinical trials are required to be conducted. Especially, in order to confirm efficacy and/or safety, it is necessary to enroll a sufficient number of patients in Phase 2b or Phase 3. Therefore, sample size calculation is one of important topics when clinical trials are designed. However, in the case of target diseases which are rare or for which enough information is not still obtained, it would be very hard to estimate the appropriate sample size at designing clinical trials. Sample size re-estimation (SSR) has been proposed by some researchers as a method to address those issues. As SSR is the method that sample size can be re-estimated during a trial by data available at that time, so it is expected that SSR would have several advantages: For example, it is not necessary to declare too large sample size when designing trials, it is expected to enroll the appropriate sample size eventually, or it is expected to keep the study power. Despite such these advantages expected, however, there are not many clinical trials that SSR is used actually. From that, practically, it is thought that there are clinical trials that should have applied SSR. FDA displayed their comments not only that SSR would cause statistical concerns in the guidance of Adaptive Designs for Clinical Trials of Drugs and Biologics (2019), but also that adaptive designs can provide a variety of advantages over non-adaptive designs. . Therefore, it is thought that SSR can be regarded as a statistical strategy to address problems which we have at designing clinical trials. In this presentation, we show the current situation of how often SSR is actually applied in clinical trials by using official database (e.g. ClinicalTrials.gov) and summarize the conditions under which applying SSR can be considered.

**AUTHORS/INSTITUTIONS:** S. Orihara, J. Moriya, K. Motoyama, Biometrics Department, R&D Division, Kyowa Kirin Co., Ltd., Chiyoda-ku, Tokyo, JAPAN|

**CONTROL ID:** 3367990

**TITLE:** Determinants of Unintended Pregnancy among Currently Married Women in Uganda

**ABSTRACT BODY:**

**Abstract Body:** Unintended pregnancy is a world health concern because of its negative association with adverse physical, social, economic, and psychological impact. This concern is no longer bound to teenagers or school going children, married women as well experience unplanned pregnancies in Uganda though little has been investigated on them. The study therefore, sought to examine the factors that influence unintended pregnancy among married women in Uganda.

The study used data from the 2016 UDHS which comprised of 11,223 married women aged 15-49 years. The data was then analyzed using frequency distribution, logistic regression, Poisson regression, log-rank test for survival functions, cox proportional hazards model, and the generalized structural equation model.

More than 44.6% of the pregnancies were unintended while 3 in 10 married women were also not using contraceptives. At the bivariate level; unintended pregnancy was significantly associated with the highest wealth quintile while contraceptive use was associated with higher education level. Similarly, children ever born were associated with married women from rural areas. At the multivariate level, married women from the northern region who were using contraceptives were 45% less likely to experience unintended pregnancy compared to their counterparts in the central. Additionally, for any additional child born among Muslim married women, the risk of unintended pregnancy raises by 4% as compared to Catholic married women.

In conclusion, married women already burdened with higher fertility were reported to suffer more from unintended pregnancy. Regional differences, religion, mass media, partner's demographic factors and improper use of contraceptive use were significantly noted to influence unintended pregnancies. The government should therefore invest in programs and policies involving sensitization of women on the effectiveness in use of contraceptives especially those in rural areas, followed by distribution of free, long-acting and quality contraceptive methods. This will help families meet their required needs and reduce on public expenditure on health.

**AUTHORS/INSTITUTIONS:** R. WASSWA, A. Kabagenyi, L. Atuhaire, Makerere University, Kampala, UGANDA|

**CONTROL ID:** 3367998

**TITLE:** Augmented Space-filling Designs for Spatial Prediction

**ABSTRACT BODY:**

**Abstract Body:** The problem of constructing spatial sampling designs in the absence of data arise from many contexts such as investigating a relatively new process/phenomena. For example, the impact of hydraulic fracturing on underground water systems or setting up an air pollution monitoring network in a place where none is existent. In such cases, space-filling designs, based on geometric criteria, are useful as they largely avoid issues concerning specifying the covariance structures required for spatial modeling. Thus, space-filling designs can be used to serve exploratory purposes then be modified accordingly once some hindsight about the process has been gained from the initial data samples. One way of modifying an initial design is by extending the initial design with a several of unsampled points. This is referred to as design augmentation. Design augmentation needs to be based on some optimal criterion which takes into account the precision of predictions at unsampled points and uncertainty of parameter estimates. Kriging variance is a measure precision of the spatial predictions while the D-criterion, the determinant of the covariance matrix, measures the uncertainty of parameter estimates. We therefore propose a compound criterion consisting of the kriging variance and D-criterion to augment the initial space-filling designs. Furthermore, we investigate the efficiency of sequential augmenting the initial design versus augmenting the initial design with a batch of points at once.

**AUTHORS/INSTITUTIONS:** K. Kebotsamang, University of Botswana, Thamaga, BOTSWANA|

**CONTROL ID:** 3368002

**TITLE:** Quantitative Interpretation of Sedia Limiting Antigen Avidity Assay for HIV Recency at Diagnosis

**ABSTRACT BODY:**

**Abstract Body:** Joseph B. Sempa<sup>1</sup>, Alex Welte<sup>1</sup>

DST-NRF Centre of Excellence in Epidemiological Modelling and Analysis, Stellenbosch University

**Background**

Laboratory based testing for recent HIV infection has become common to estimate HIV incidence from cross-sectional surveys. Individual-level uses are less clear, but numerous studies and pilots of diagnostic services have been returning results of recency tests to individuals. There is considerable interest in using such information to support posttest counselling, disclosure decisions, and contact tracing. The formal information content of individual recency test results, as pertaining to estimation of timing of infection, is disputed.

**Methods**

We calibrated a simplistic 'anti-HIV antibody avidity' biomarker progression model (exponential approach to subject-specific asymptote) for the Sedia LAg avidity Elisa assay, to a previously described rich data set from the CEPHIA collaboration. This effectively provides an approximate likelihood function describing the distribution of assay results as a function of time since seroconversion. We considered possible assay values obtained at a first HIV positive test, assuming a uniform prior on seroconversion time since the last negative test, and calculated the infection time posterior.

**Results**

The calibrated biomarker progression model provides a family of percentile 'growth curves' like weight for age curves etc. We can vary inter-test intervals and hypothetical avidity marker values obtained at the first positive test date. Within the usable dynamic range of the assay, we can compare priors and posteriors for a range of inter-test intervals and identify situations in which the avidity marker provides substantial additional information, suggesting that seroconversion took place significantly early or late between the last negative and first positive test.

**Conclusion**

Infection time estimates for frequent testers cannot be substantially improved with additional recency testing, but for infrequent testers, it is clear that these estimates are substantially improved through the use of quantitative recency testing, rather than just relying on testing history. Such timing information would be expected to strengthen psychosocial aspects of post-diagnosis counselling and management as have already been identified as being of interest, and it is also relevant to studies of early post-infection biology.

**AUTHORS/INSTITUTIONS:** J.B. Sempa, A. Welte, DST-NRF Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA), University of Stellenbosch, Stellenbosch, South Africa, Stellenbosch University, Stellenbosch, Western Cape, SOUTH AFRICA|

**CONTROL ID:** 3368010

**TITLE:** Combining prediction models from meta-analytical perspective to maximize AUC

**ABSTRACT BODY:**

**Abstract Body:** Development of accurate clinical prediction models has been main focus in the medical literature and the number of published studies that share the same prediction task or population have been increased substantially over the past few decades. In the context of meta-analysis, previous studies have mainly aimed to synthesize regression coefficients of binary prediction models from the estimation perspective: the synthesis is designed to improve the estimation accuracy of the pooled effect. In contrast, our study focuses on synthesizing the regression coefficients from the prediction perspective: the synthesis is designed to maximize prediction accuracy of the synthesized model that is measured by the area under the receiver operating characteristic curve on one available dataset. Here, we are interested in a problem of prediction of a binary outcome by utilizing the information for summary statistics reported from  $K$  past studies which share the similar objective to the current study with the same outcome and a (sub)set of covariate vectors. This study proposes a new linear predictor for the outcome integrating these information under a distributional assumption between the current study and  $K$  studies in a context of meta-analysis.

**AUTHORS/INSTITUTIONS:** D. Yoneoka, St. Luke's International University, Tokyo, Chuo-Ku, JAPAN|

**CONTROL ID:** 3368023

**TITLE:** Bayesian model averaging highlights dependences of clonal dynamics on different lentiviral vectors in Gene Therapy studies

**ABSTRACT BODY:**

**Abstract Body:** In gene therapy, vector integration sites (IS) are a proxy for clonal identity and allow the assessment of safety and efficacy of the treatment. Indeed IS data allow to describe how the clonal population is composed and evolve over time. How the genotype of the treated subject and the lentiviral vector design used may affect the aforementioned readouts is still not well understood. Indeed, we retrieved IS data longitudinally from a mice study. A group of tumor prone mice and control mice were treated either with a neutral or with a toxic vector[1].

We analyze the behaviour of the clonal diversity, measured via the log-Shannon Entropy, over time as a function of the viral vector and the mouse genotype in each cell lineage. We also propose rescaling variables as potential covariates such as the sample DNA mass, the pool size and the vector copy number. Indeed those variables could affect the cellular counts and in turn their shannon entropy. We then consider a set of Bayesian regression models featuring different combinations of the covariates. We select the most likely features according to the marginal likelihood and we average across them. We compare Laplace Integration and Thermodynamic Integration methods with Information Criteria to estimate the marginal likelihood in a simulation study. According to the simulations Laplace Integration outperforms all the remaining methods and we therefore use it on the real data.

As a result the trends of the entropy decay is the same for all lineages. Furthermore, we found that the genotype has no effect on the entropy decay either. Indeed our regression framework revealed that the quicker decay for the tumour prone genotype is only caused by the higher mortality of mice of that group, yielding a systematically faster decay of the entropies. Unlike the genotype effect, the vector effect stays significant after adjusting for potential confounding variables. This is an important clinical finding, as the clinicians have to choose the vector for a gene therapy treatment. Therefore, our experimental framework allows to describe the impact of the genomic integration of different vector designs on the behavior of thousands of cells in different cell lineages, in normal and tumor-prone genetic backgrounds.

References

[1] Asgct 21st annual meeting abstracts. Molecular Therapy, 26(5, Supplement 1):1-459, 2018.

**AUTHORS/INSTITUTIONS:** L. Del Core, M. Grzegorzcyk, Bernoulli Institute, University of Groningen, Groningen, NETHERLANDS|L. Del Core, A. Calabria, D. Cesana, E. Montini, San Raffaele Telethon Institute for Gene Therapy (SR-Tiget), Milan, ITALY|E. Wit, Institute of Computational Science, Università della Svizzera Italiana, Lugano, SWITZERLAND|

**CONTROL ID:** 3368025

**TITLE:** Offline and online control of the scaled false discovery rate.

**ABSTRACT BODY:**

**Abstract Body:** In multiple hypotheses testing, safeguard against type I errors is usually the primary concern, according to the Neyman-Pearson principle. Typically, type I errors are controlled inherently through an error rate that involves the number of false positives (FP).  $FWER=P(FP>0)$  and  $FDR =E(V/s(R))$ , where  $R$  is the number of rejections, are the commonly used error rates in practice. We present the scaled false discovery rate  $sFDR=E(FP/s(R))$ , where  $s$  is a non-decreasing function. The  $sFDR$  can be regarded as an extension of the  $FDR$ , in the sense that the influence of the number of rejections in the control of type I errors is governed by the scaling function  $s$ . Controlling the  $sFDR$  when  $s(R)=1$  guarantees  $FWER$  control, while  $FDR$  is the particular case corresponding to  $s(R)=R$  (identity function). We consider two types of control. First, the offline control, which corresponds to the situations where all  $p$ -values corresponding to the hypotheses tested, are all available at the time of decision. Second, the online control, which corresponds to the situation where test  $p$ -values are provided one by one, and a decision has to be made after receiving each individual  $p$ -value. We prove that the offline  $sFDR$  control can be achieved via a step-up procedure with a threshold sequence proportional to  $s$ ; that is, it has  $s$  as a shape function. We also show that online control can be achieved by a procedure that generates a rejection pay-off proportional to the function  $s$ .

**AUTHORS/INSTITUTIONS:** D. Meskaldji, S. Morgenthaler, Institute of Mathematics, EPFL, Lausanne, SWITZERLAND|

**CONTROL ID:** 3368028

**TITLE:** On the Statistical ranking of the Bioinformatics Software on Free and Open Source Software Community

**ABSTRACT BODY:**

**Abstract Body:** One major advantage of medical technology is the Predict-and-Prevent model in health care system. Bioinformatics software, being one of these technologies, could be faced with the high cost, in the case of proprietary software, hence, reducing the accessibility and the usage level of bioinformatics software technology in general. With the revolution of Free and Open Source (FOS) software, there is need to continuously assess the contribution of FOS medical software to the health care industry. This study had ranked the FOS bioinformatics software among other FOS Software on sourceforge online FOS community, based on usage level and the download size. Two-Phase sampling for regression estimation method in Survey Statistics was used to obtain the total software raters, the associated Mean Square Errors (MSEs) and the percentage coefficient of variation. The MSE and the coefficient of variation were used as the yardsticks for ranking of the FOS software. It was discovered that FOS bioinformatics software was ranked fifth based on the estimated total raters (with the estimated total of 61532 total software users) while it was ranked as the seventh, among the FOS software, based on the computed percentage coefficient of variation (with the percentage estimated coefficient of variation of 12.27%). Recommendations on the wide usage of Bioinformatics FOS software were documented for the FOS software developers and the users.

**Keywords:** Bioinformatics software, free and open source, sourceforge, software users, download

**AUTHORS/INSTITUTIONS:** P.I. Ogunyinka, D.A. Agunbiade, Mathematical Sciences, Olabisi Onabanjo University, Ijebu-Ide, Ogun, NIGERIA|E.F. Ologunleko, Statistics, University of Ibadan, Ibadan, Oyo, NIGERIA|P.I. Ogunyinka, Information Technology, Apple Microfinance Bank, Ijebu-Ode, Ogun, NIGERIA|

**CONTROL ID:** 3368037

**TITLE:** Evaluation of dynamic prediction models in a Hidden Markov setting

**ABSTRACT BODY:**

**Abstract Body:** Different models to dynamically predict survival in the presence of a longitudinal biomarker have been proposed in the literature. While joint modeling of longitudinal and time-to-event data was generally found to outperform landmarking in previous simulation studies, the data for these studies were generated from a joint model. To determine whether the previous findings could be generalized to a setting where the data were generated from a Hidden Markov model (HMM), we performed a simulation study. We first fitted an illness-death HMM to the data from an observational cohort of 475 acute heart failure patients. We subsequently used this model to repeatedly generate training and test datasets. Three different approaches were considered to model the conditional failure probabilities at different landmark times and window widths in the training datasets: landmarking, joint modeling, and Hidden Markov modeling. For the HMM, the Viterbi algorithm was applied to predict the hidden state at the selected landmark time in the test datasets. For the joint model, shared random-effect models with random intercept and random slope for the longitudinal model were considered. For landmarking, both naive landmarking and two-step landmarking were considered. The predictive performance of the three approaches were compared in terms of discrimination (AUC) and calibration (Brier score). We found that the HMM performed better than the other two models in terms of Brier score, especially when longer window widths were selected. Surprisingly, the joint model was still robust in the HMM setting.

**AUTHORS/INSTITUTIONS:** Y. Chen, D. Postmus, University of Groningen, Groningen, NETHERLANDS|

**CONTROL ID:** 3368045

**TITLE:** Predictions by random forests - confidence intervals and their coverage probabilities

**ABSTRACT BODY:**

**Abstract Body:** Random forests are a popular supervised learning method. Their main purpose is the robust prediction of an outcome based on a learned set of rules. To evaluate the precision of predictions their scattering and distributions are important. In order to quantify this, 95% confidence intervals for the predictions can be generated using suitable variance estimators. However, these variance estimators may be under- or overestimated and the confidence intervals thus cover ranges either too small or too large, which can be evaluated by estimating coverage probabilities through simulations. Therefore, the aim of our study was to examine coverage probabilities for two popular variance estimators for predictions made by random forests, the infinitesimal jackknife according to Wager et al. (2014) and the fixed-point based variance estimator according to Mentch and Hooker (2016). We performed a simulation study considering different scenarios with varying sample sizes and various signal-to-noise ratios. Our results show that the coverage probabilities based on the infinitesimal jackknife are lower than the desired 95% for small data sets and small random forests. On the other hand, the variance estimator according to Mentch and Hooker (2016) leads to overestimated coverage probabilities. However, a growing number of trees yields decreasing coverage probabilities for both methods. A similar behaviour was observed when using real datasets, where the composition of the data and the number of trees influence the coverage probabilities. In conclusion, we observed that the relative performance of one variance estimation method over the other depends on the hyperparameters used for training the random forest. Likewise, the coverage probabilities can be used to evaluate how well the hyperparameters were chosen and whether the data set requires more pre-processing.

**AUTHORS/INSTITUTIONS:** D. Kormilez, B. Laabs, I.R. König, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany, Lübeck, GERMANY|

**CONTROL ID:** 3368046

**TITLE:** Quality control in genome-wide association studies revisited:  
a critical evaluation of the standard methods

**ABSTRACT BODY:**

**Abstract Body:** Genome-wide association studies (GWAs) investigating the relationship between millions of genetic markers and a clinically relevant phenotype were originally based on the common disease - common variant assumption (Ziegler, König & Pahlke, 2012), thus aiming at identifying a small number of common genetic loci as cause for common diseases. Given the enormous cost reduction in the acquisition of genomic data, it is not surprising that since the first known GWA by Klein et al. (2005), this study type was established as a standard method. However, since even low error frequencies can distort association results, extensive and accurate quality control of the given data is mandatory. In recent years, the focus of GWAs has shifted (Visscher et al. 2017), and the task is no longer primarily the discovery of common genetic loci. Also, with increasing sample sizes and (mega-)meta-analyses of GWAs, it is hoped that loci with small effects can be identified. Furthermore, it has become popular to aggregate all genomic information, even loci with very small effects and frequencies, into genetic risk prediction scores, thus increasing the requirement for high-quality genetic data.

However, after extensive discussions about standards for quality control in GWAs in the early years (Ziegler et al. 2008), further work on how to control data quality and adapt data cleaning to new GWAs aims has become scarce. The aim of this study therefore was to perform an extensive literature review to evaluate currently applied quality control criteria and their justification.

Our results show that in most published GWAs, no scientific reasons for the applied quality steps are given. Cutoffs for the most common quality measures are mostly not explained. Especially the principal component analysis and the test for deviation from Hardy-Weinberg equilibrium are frequently used as quality criteria in many GWAs without analyzing the existing conditions exactly and adjusting the quality control accordingly.

Building on the findings from the literature search, a workflow was developed to include justified quality control steps, keeping in mind that a strict quality control, which removes all data with a high risk of bias, always carries the risk that the remaining data is too homogeneous to make small effects visible. This workflow is subsequently illustrated using a real data set.

**AUTHORS/INSTITUTIONS:** H. Bruderemann, T.K. Rausch, I.R. König, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, GERMANY|T.K. Rausch, Klinik für Kinder- und Jugendmedizin, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, GERMANY|

**CONTROL ID:** 3368049

**TITLE:** Finding Simple and Optimal Architecture and the Effect of Skip Connections in Fully Convolutional Networks (FCN) for MRI-Based Brain Tumor Segmentation

**ABSTRACT BODY:**

**Abstract Body:** Fully Convolutional Networks (FCN) such as U-Net is a strong deep learning technique for MRI-based brain tumor segmentation but required a lot of computational resources due to its complexity of architecture. In this study, we aim to find the simple and optimal architecture of FCN for our specific MRI dataset. We experiment with the networks by adding the skip connections in FCN since skip connections are used to recover spatial information lost during downsampling. We also study the effect of skip connections in FCN to performance of segmentation, we imitate the skip connections in U-Net since U-Net uses skip connection in its architecture. Because of segmentation task is similar to the classification of imbalanced class for every pixel in the image, we use weighted binary cross-entropy for the loss function of our networks. For evaluating the performance of the networks, we use mean intersection over union (mIOU) which means of the ratio of the overlapping area of ground truth and predicted segmentation area to the total area.

**AUTHORS/INSTITUTIONS:** N. Iriawan, M. Almuhyar, Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, East Java, INDONESIA|A.A. Pravitarsi, Statistics, Universitas Padjadjaran, Bandung, West Java, INDONESIA|

**CONTROL ID:** 3368052

**TITLE:** A Sparse Partial Least Squares (PLS) Regression

**ABSTRACT BODY:**

**Abstract Body:**

With the recent advancements in technology it is common to measure several outcomes and predictor variables about a patient. This causes collinearity in the data, making standard regression techniques ill-suited for modelling. To handle collinearity, we use PLS and select active variables for better interpretation by introducing a penalty function to the objective function.

Our work is motivated by data on patients suffering from scleroderma, we are interested in identifying biomarkers that are correlated with skin hardening (mRSS), and abnormal lung functions (DLCO, FVC). There are 93 patients, 50 biomarkers and 3 outcomes. Identifying active biomarkers will lead to the measurement of fewer biomarkers in the future.

Interest is on the joint modelling of the three outcomes. Some state-of-the-art methods as well as envelope model of Cook et. al. (2013) are used for reducing the dimension of data but cannot select active.

A new sparse envelope estimator is proposed for joint variable selection. Sparsity improves statistical models because it makes variable selection and interpretation possible. The proposed method uses a penalized normal likelihood formulation by shrinking estimators towards zero via a lasso penalty function. The proposed method is compared to SPLS (Chun & Keles, 2010) and E-SPLS (Zhu & Su 2019) using simulated and real data. Further, variable-selection is performed for the outcome variables together.

**AUTHORS/INSTITUTIONS:** E.C. Uzochukwu, J. Houwing-Duistermaat, J.T. Kent, Department of Statistics, University of Leeds, Leeds, UNITED KINGDOM|

**CONTROL ID:** 3368056

**TITLE:** Zero Truncated Negative Binomial with Excess Ones

**ABSTRACT BODY:**

**Abstract Body:** Zero truncated negative binomial is modified to include excess ones so as to improve goodness of fit. This is necessary when data is dispersed and zero has been eliminated from data. However, when ones are unduly large, the proportion of this excess has to be estimated to improve the fit. This development is applied using a national survey data.

**AUTHORS/INSTITUTIONS:** E.T. Jolayemi, Department of Statistics, University of Ilorin, Ilorin, Kwara, NIGERIA|

**CONTROL ID:** 3368057

**TITLE:** Integrating Biological Knowledge and Omics Data Using Pathway Guided Random Forests: A Benchmarking Study

**ABSTRACT BODY:**

**Abstract Body:** High-throughput technologies allow comprehensive characterization of individuals on many molecular levels. However, training prediction models on omics data is challenging. A promising solution is the integration of external knowledge about structural and functional relationships into the modeling process. We compared four published random forest based approaches using two simulation studies and nine experimental data sets. In the first simulation study with many associated pathways synthetic features (SF) and prediction error (PE) showed high empirical power across different pathway sizes, degrees of association and correlation patterns, whereas Hunting and Learner of Functional Enrichment (LeFE) were only able to detect large pathways with strong signal. In the complete null scenario, SF showed increased false discovery rates for all pathways. In the second simulation study with a single associated pathway and realistic correlation patterns all methods had similar power, but as in the other simulation study, SF was the only method with elevated false discovery rates. In the experimental data sets PE and SF usually identified the target pathway but additionally detected almost all other pathways. Hunting and LeFE had lower detection frequencies but rarely selected additional pathways. In conclusion, the self-sufficient PE approach should be applied when large numbers of relevant pathways are expected. The competing methods Hunting and LeFE should be used when low numbers of relevant pathways are expected or the most strongly associated pathways are of interest. The hybrid approach SF is not recommended because of its high false discovery rate.

An R package providing functions for data analysis and simulation is available at GitHub (<https://github.com/szymczak-lab/PathwayGuidedRF>).

**AUTHORS/INSTITUTIONS:** S. Szymczak, S. Seifert, Institute of Medical Informatics and Statistics, Kiel University, Kiel, GERMANY|

**CONTROL ID:** 3368097

**TITLE:** How Biomedical Statisticians Can Keep Abreast of High-Impact Statistical Papers

Michael J. Schell, Moffitt Cancer Center, Tampa, FL, U.S.A.

Ji-Hyun Lee, University of Florida, Gainesville, FL, U.S.A.

**ABSTRACT BODY:**

**Abstract Body:** We have identified a total of 3655 high-impact statistical papers from the Web of Science by high citation counts, with 100 overall and  $\geq 20$  in 2014-15 or 2015-2016, or  $\geq 40$  in 2015-2016. This represents at most 3% of all statistical papers written in major statistics relevant peer-reviewed journals since 1900. This presentation provides an overview of these papers, with the aim of providing statistical practice tools for applied biomedical statisticians. A 4-level hierarchical classification system has been developed for the papers, with 6 kingdoms: Specialized (28%), Regression (25%), Estimation (16%), Experimental studies (15%), Broad topics (10%), Broad testing (7%). The kingdoms are divided into 51 phyla, and then 287 classes. Twenty phyla are of greater interest to biomedical statisticians: comprising 873 of the 1163 papers with medicine, epidemiology or social science (principally psychology) as their highest citation source, including 6 phyla in experimental studies and 5 phyla in regression. The top 5 phyla are: Meta-analysis (N=112), Survival (N=106), Epidemiology (N=86), Clinical trials design (N=63) and Clinical trials analysis (N=49). Phyla with the most recent new research are (median publication year in parenthesis): Missing data (2005), Causal models (2002), Diagnostic testing (2001), and Clinical trials analysis (2000). Phyla with the greatest median diffusion (ratio of total citations to the citations from the top citing Web of Science journal category) across research arenas are: Structural equations modeling (6.1), Correlation (5.3), Broad testing/Categorical (5.1), Meta-analysis (5.1), Diagnostic testing (4.8), and Factor analysis (4.8). Besides the papers we have identified, surveillance of 7 journals is recommended for identifying new research tools; Statistics in Medicine (N=66), American Journal of Epidemiology (22), STATA Journal (18), Journal of Statistical Software (15), Epidemiology (15), Journal of Clinical Epidemiology (10), and Statistical Methods in Medical Research (10), with the number of articles since 2005 among the 3655 given in parentheses.

**AUTHORS/INSTITUTIONS:** M.J. Schell, Biostatistics and Bioinformatics, Moffitt Cancer Center & Research Institute, Tampa, Florida, UNITED STATES|J. Lee, Biostatistics, University of Florida, Gainesville, Florida, UNITED STATES|

**CONTROL ID:** 3368135

**TITLE:** METHODOLOGICAL IMPLEMENTATION OF HEIGHT HERITABILITY ESTIMATION IN GENOME ASSOCIATION STUDIES.

**ABSTRACT BODY:**

**Abstract Body:** In Genome-Wide Association Studies, height is a polygenic trait that suffers from the Missing Heritability Problem.; a condition where quantitative traits with high heritability show very low values of SNP-heritability. In this study, we investigate two statistical models used in polygenic analysis namely E-Bayes and Efficient Mixed Model Association (EMMA) under the Linear Mixed Model framework. We evaluate the expected sensitivity of these models to inherent heritability in order to identify the most appropriate model for the estimation of height heritability.

**AUTHORS/INSTITUTIONS:** O.N. Ezichi, Statistics, University of Ibadan, Nigeria., Ibadan, Oyo, NIGERIA|

**CONTROL ID:** 3368141

**TITLE:** Evaluating study-level surrogacy of binary outcomes using a novel Bayesian bivariate meta-analytic approach

**ABSTRACT BODY:**

**Abstract Body:** Surrogate endpoints play an important role in drug development when they can be used to measure treatment effect early compared to the final clinical outcome and to predict clinical benefit or harm. Such endpoints are assessed for their predictive value of clinical benefit by investigating the surrogacy between treatment effects on the surrogate and final outcomes using bivariate meta-analytic methods.

The standard meta-analytic approach models the observed treatment effects on the surrogate and final outcomes jointly, at both the within-study and between-studies levels, using a bivariate normal distribution. For binomial data a normal approximation can be used on log odds ratio scale, however, this method may lead to biased results when the proportions of events are close to one or zero, affecting the validation of surrogate endpoints. In this paper, two alternative Bayesian meta-analytic approaches are introduced which allow for modelling the within-study variability using binomial data directly. The first method uses independent binomial likelihoods to model the within-study level avoiding to approximate the observed treatment effects. This method, however, ignores the within-study association. The second method we developed, models the summarised events in each arm jointly using a bivariate copula with binomial marginals. This allows the model to take into account the within-study association through the copula dependence parameter. We carried out a simulation study to assess the performance of the proposed methods and compare them with the standard bivariate normal model in 18 data scenarios.

We also applied the methods to an illustrative example in chronic myeloid leukemia.

**AUTHORS/INSTITUTIONS:** A. Papanikos, K. Abrams, J. Thompson, S. Bujkiewicz, Health Sciences, University of Leicester, Leicester, UNITED KINGDOM|

**CONTROL ID:** 3368150

**TITLE:** A Bayesian state space treatment transition framework for modelling HIV adolescent and young people Patients in Zambia

**ABSTRACT BODY:**

**Abstract Body:** Modelling treatment outcome for data coming from real world setting, presents numerous fundamental challenges. Patient in care experience several intermediate treatment events before experiencing the terminal outcome. Intermediate states creates dependence between several other states. We developed a three state space framework and applied a semi markov transition model to characterize adolescent and young patient-level dynamic treatment trends. Unobserved effect specific susceptibility-was handled via a random effects model. We hypothesize that adolescent and young people are enrolled on treatment with the probability of  $p_{01}$ ; transition to second line with the probability of  $p_{12}$  and die with the probability of  $p_{23}$ . Probability from first line to death is  $p_{13}$ .

Bayesian multi state cox model was fitted using non informative priors. Gaussian priors were assigned to regression parameters with zero mean and wider variance of 1000. A weakly prior density with gamma (0.01, 0.01) was assigned to transition rates and precision parameters, Gaussian priors were given to frailty terms ( $N(0, t^2)$ ). INLA, was used to generate the marginal posterior distributions of the various transition parameters.

**Results:** Out of the 13,266 adolescent clients enrolled, 10770(78.8%) remained on the first line treatment, 1494(11.3%) transitioned to second line treatment and 962 (7.3%) died on the first line treatment. Overall, staying on the first line treatment was higher compared to staying on the second line ART treatment. Probability of transitioning to the second line treatment was  $p=0.07$ , while that of remaining on the first line was 0.935, and on second line treatment was 0.997. Probability of transition to second line from first line was determined to be 0.029. Significant heterogeneity was observed at patient and province for those from first line to second line; patient and facility for those from first line to death; from second line to death, heterogeneity was at patient, facility and district level.

**Implications:** The evidence of rampant delays on first line and low probability of transitioning to second line treatment has documented laps in treatment guideline enforcement at facility levels. Being on first line treatment for long time, reveals unresponsive surveillance system that fails to detect failure in treatment regimen.

**AUTHORS/INSTITUTIONS:** I. Fwemba, Biostatistics, University of Ghana, Accra, GHANA|I. Fwemba, Department of Epidemiology, University of Zambia, Lusaka, Lusaka, ZAMBIA|

**CONTROL ID:** 3368160

**TITLE:** Mixed effects model to assess the risk factors associated with change in child Body mass index in South Africa

**ABSTRACT BODY:**

**Abstract Body:**

Child obesity is a global public health concern and it is an area, which needs focus in order to improve the nation's health. Obesity is associated with significant health risks and comorbidities such as cardiovascular disease and type 2 diabetes. Child obesity usually defined by body mass index (BMI) for children aged two years or older and by weight-for-length for children younger than two years. The aim of this study is to identify significant determinants of the change in child BMI focusing in children age 5 to 17 years using the data from the National Income Dynamic Study (NIDS). The exploratory data analysis shows that the BMI does not remain constant, or follow a linear pattern overtime. Therefore, to assess changes in the BMI while accounting for the correlation between the repeated measurements of each child and the multilevel structure of the data, we utilized the framework of linear mixed effects models. In the fixed effects part, beside demographic, socioeconomic and other factors such as child caregiver information and their backgrounds on child health, we allowed for a nonlinear effect of time using natural cubic splines with two internal knots placed at the corresponding percentiles of the time points. In the random effects structure, we included random intercepts and random nonlinear splines using the same splines as in the fixed effects part. The overall results show that gender, age and race of a child, relationship of person responsible for care of a child, education level of person responsible for care of a child, household income and poverty level of a household were the potential determinants for change of child BMI overtime. These findings could have important implications in the public health policy.

**AUTHORS/INSTITUTIONS:** L. Debusho, Statistics, University of South Africa, Florida, Johannesburg, Gauteng, SOUTH AFRICA|

**CONTROL ID:** 3368176

**TITLE:** Block design with nested rows and columns for plant protection research

**ABSTRACT BODY:**

**Abstract Body:** The specific character of research on plant protection implies the necessity to studies on planning and analysis of such experiments (e.g. Kozłowska et al. 2014). Plant protection experiments are designed on heterogeneous experimental material, for example, in the case of non-uniform or particularly low levels of disease or pest infection. There are frequently factorial or near-factorial experiments. In such experiments the importance of interaction and hidden replication are emphasized.

Block design with nested rows and columns is frequently used. A design is said to have nested rows and columns if the set of experimental units is partitioned into blocks and each block is further partitioned into rows and columns. Thus it is reasonable to seek a design that can withstand the loss of blocks. Several authors have investigated conditions for robustness of some block designs (e.g. Godolphin and Godolphin 2015). The robustness of block design with nested rows and columns in the face of loss of whole blocks are presented.

The theory of optimal experimental designs is concerned with the problem of selecting a design which minimizes some functional of the information matrix over all possible designs in considered class. The commonly used criteria are called A, D and E criteria. Bagchi and Bagchi (2001) introduced the definition that a design  $d$  is said to be better than another design  $d'$  from the some class in the sense of majorization. We extend the theory of block designs with nested rows and columns. We study optimality of the designs possessing special property with respect to any criterion of a described form. For any design described by mixed model of observation we consider the optimality of the design, particularly when only the bottom stratum is used.

Examples of block designs with nested rows and columns applied to the plant protection experiments mentioned above are described.

Bagchi B, Bagchi S. (2001). Optimality of Partial Geometric Designs. *The Annals of Statistics* 29, 577–594.

Godolphin JD, Godolphin EJ. (2015). The robustness of resolvable block designs against the loss of whole blocks or replicates. *J. Statist. Planning Inf.* 163, 34-42.

Kozłowska M, Jaskulska M, Lacka A, Kozłowski RJ. (2014). Analysis of studies of the effectiveness of a biological method of protection for organic crops. *Biometrical Letters* 51, 45-56.

**AUTHORS/INSTITUTIONS:** M. Kozłowska, Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Poznan, POLAND|

**CONTROL ID:** 3368180

**TITLE:** WEIGHING DESIGNS IN AGRICULTURAL EXPERIMENTS: THEORY AND APPLICATIONS

**ABSTRACT BODY:**

**Abstract Body:** The problematic aspects of this presentation are issues linked to the planning of the experiments. In the theory of experiments taking into account planning of the experiments and, afterwards, analysing and the formulation of conclusions, a significant role-plays the way of the experiment planning. It is required the information of experimental material, the environments conditions are carried out in tandem with optimal statistical properties of the considered designs. It is needed to maintain the balance between these expectations to gain the most effective experiment. Both the increase of the effectiveness and the reduction of the error variance depend on taking appropriate assumptions connected with experimental errors and the design. Many reasons influence on the value of the variance: some inaccuracy in measuring of the examined characteristic, in agricultural efforts, non-homogeneity of experimental designs (for example non-homogenous field richness). Hence, we need to plan and perform (carry out) the experiment in such a way to hold down the influence of these reasons having regard to the aim and the manner of performing the experiment. We consider the experiment in that the random vector of observations equals the product of the design matrix and the vector of unknown measurements. We study this experiment under different assumptions regard to the vector of errors. A target point in my research is such planning of the experiments in the sense of attaining the best estimators of unknown measurements of objects. Optimal designs permit to the value of unknown parameters with minimal variance and practically allow reducing the experimental costs.

We present recent accomplishments in the optimal weighing designs. We give the conditions determining optimal design and examples of application of such designs in agricultural experiments.

**AUTHORS/INSTITUTIONS:** M. Graczyk, B. Ceranka, Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Poznan, POLAND]

**CONTROL ID:** 3368294

**TITLE:** Firth adjusted score function for monotone likelihood in the mixture cure fraction model

**ABSTRACT BODY:**

**Abstract Body:** Models for situations where some individuals are long-term survivors, immune or non-susceptible to the event of interest, are extensively studied in biomedical research. Fitting a regression can be problematic in situations involving small sample sizes with high censoring rate, since the maximum likelihood estimates of some coefficients may be infinity. This phenomenon is called monotone likelihood, and it occurs in the presence of many categorical covariates, especially when one covariate level is not associated with any failure (in survival analysis) or when a categorical covariate perfectly predicts a binary response (in the logistic regression). A well known solution is an adaptation of the Firth method, originally created to reduce the estimation bias. The method provides a finite estimate by penalizing the likelihood function. Bias correction in the mixture cure model is a topic rarely discussed in the literature and it configures a central contribution of this work. In order to handle this point in such context, we propose to derive the adjusted score function based on the Firth method. An extensive Monte Carlo simulation study indicates good inference performance for the penalized maximum likelihood estimates. The analysis is illustrated through a real application involving patients with melanoma assisted at the Hospital das Clínicas/UFMG in Brazil. This is a relatively novel data set affected by the monotone likelihood issue and containing cured individuals.

**AUTHORS/INSTITUTIONS:** E.A. Colosimo, F. Almeida, V. Mayrink, Statistics, UFMG, Belo Horizonte, MG, BRAZIL|

**CONTROL ID:** 3368297

**TITLE:** Extending generalized canonical correlation analysis to family data.

**ABSTRACT BODY:**

**Abstract Body:** Recent technological advances have allowed the generation of large-scale data, such as omic data applied to several types of studies in genetics. Currently, the Big Data revolution is present in areas such as health, marketing and finances. Regarding health area, new data collection technologies allow a more in-depth study of molecular biology at genomic, transcriptomic, and proteomic levels, leading to more targeted and personalized patient healthcare solutions, refereed as precision medicine.

The multi-omics approach is revolutionary as it gathers information from multiple omics evaluated on the same individual, and has a great impact on the methods of analysing in this kind data. Integration methods for these databases have been proposed in the literature, based mainly on matrix factorization in terms of individual scores and variable loadings. Data integration refers to the situation where, for a given system, multiple data sources are available and we aim to study them in an integrated manner to improve knowledge.

However, these techniques assume that the sample is composed of independent individuals, an assumption that is not satisfied for family studies, commonly used in genetics and involving individuals related by kinship. Family data not only include genetic information throughout the genome, but also contain correlation among individuals.

Thus, we focus on developing new statistical methodologies appropriate to the analysis and integration of multi-omics databases in structured family designs that allow classification of individuals and construction of biomarkers. The main idea is to extend generalized canonical correlation analysis (GCCA) taking into account the family correlation between individuals and apply these methodologies to simulated data and real data, such as data from cohort studies including Brazilian families.

GCCA is a powerful tool used in database integration, allowing the connection of more than two data sets. The development of complementary analysis of this approach, such as its generalization to family data situations, leads to a better understanding of the relationship among multi-omics data and human phenotypes. By using simulation studies we show the flexibility of our extended method in finding data integration under decomposition of polygenic and error components.

Keywords: Family-based design, GCCA, Multi-omics data.

**AUTHORS/INSTITUTIONS:** C.M. Tondolo, J.M. Pavan Soler, University of Sao Paulo, São Paulo, Sao Paulo, BRAZIL|

**CONTROL ID:** 3368332

**TITLE:** Application of Randomized Response Technique on Undergraduate Students Sexual Abuse Sensitive Questions in Nigerian University.

**ABSTRACT BODY:**

**Abstract Body:** Students sexual abuse represents one form of organizational wrong doing in the context of an academic organization. Previous research has been hampered by a paucity of accurate data regarding the validity of survey responses estimating the occurrence of college students sexual abuse because the researchers are frequently confronted with the problem of extracting accurate data on sensitive matters from the respondents during the conduct of any socio economic and the related surveys. For sensitive questions, the respondents are regularly reluctant to divulge genuine information. A revolutionary technique, Randomized Response Technique (RRT), is specifically designed approach to enhance the accuracy of responses to sensitive questions. Through this approach, interviewer is able to estimate the proportion of respondents associated with the 'sensitive question' without knowing the actual status of individual respondent. RRT has been used substantially in some countries to accumulate data on sensitive issues like the proportion of tax evaders in the country. However, RRT is sparingly use in Nigeria. This could be efficiently be employed to numerous field like health, economics and problems related to social stigmas such as sexual abuse etc. Students sexual abuse is not only violation of their rights but it could additionally cause them to be psychologically depressed and even place them in long-term psychological trauma. It is really difficult to get accurate responses due to danger and stigma attached to it. RR surveys were conducted on initial and replication samples of Mass Communication students (total n = 312) and 84 Business Administration under graduate botany students. The estimates of prevalence of students sexual abuse victimization were 22.7 % in the initial and 17.7 % in the replication samples, with the difference not significant ( $z = 1.25, y > .05$ ). The findings are discussed within the context of results from conventional surveys.

Keywords: Randomized response technique, Students sexual abuse.

**AUTHORS/INSTITUTIONS:** E.F. Ologunleko, P.I. Ogunyinka, A.A. Sodipo, Department of Statistics, University of Ibadan, Ado Ekiti, Ekiti, NIGERIA|

**CONTROL ID:** 3368348

**TITLE:** breedingSimulatR: An R package to simulate breeding schemes.

**ABSTRACT BODY:**

**Abstract Body:** Rapid advancement of genotyping technologies allows access to a large amount of whole-genome marker data, which enables genomic selection (selection based on the prediction of genetic values from whole-genome marker genotypes) to improve genetic gain for developing new varieties. Various schemes for the implementation of genomic selection have been developed in order to enhance the genetic gain under different constraints, especially for short-term and long-term genetic gain. It is, however, difficult to validate the potential of these schemes empirically, because of the large amount of financial, material, and temporal resources required to conduct “real” breeding programs. However, this limitation can be overcome by simulating breeding programs. From this perspective, we are developing breedingSimulatR, which is an R package providing tools to develop such simulation algorithms. Using the package, a user is able easily (1) to define the species characteristics, such as number of chromosomes, chromosomes length, recombination rate, (2) to specify information about Single Nucleotide Polymorphism (SNPs) markers, (3) to set-up an initial population based on real or simulated genetic data, (4) to define a genetic architecture for one or several traits, (5) to define functions to select and mate individuals, (6) to run thousands of breeding campaigns, (7) to get several statistics about the simulated breeding campaigns: distribution of the final genetic gain of the population, distribution of the breeding value of the final best individual, distribution of the genetic diversity for each generation. The analysis of these statistics provides to the user important information for the assessment of the tested schemes of genomic selection breeding and allow him/her to compare it to others. Moreover, the package will be useful as a “benchmark test” system that allows the validation of a newly proposed scheme and gives a standardized measure for the potential of the scheme.

Julien Diot, Hiroyoshi Iwata,

Graduate School of Agricultural and Life Sciences, The University of Tokyo

**AUTHORS/INSTITUTIONS:** J. Diot, Agricultural and biological science, The University of Tokyo, Toshima-Ku, Toshima-ku, JAPAN|H. Iwata, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, JAPAN|

**CONTROL ID:** 3368365

**TITLE:** A machine learning based tandem approach to predict extubation failure in pediatric ICU.

**ABSTRACT BODY:**

**Abstract Body:** Pediatric cardiac critical care providers are often challenged with the equally important but often conflicting goals of minimizing patients' exposure to mechanical ventilation and preventing extubation failure. Extubation failures have been associated with adverse outcomes including cardiac arrest and mortality. Reliable measures of extubation readiness, while validated in adult patients, remain elusive in pediatric cardiac critical care. Patients in the cardiac intensive care unit (CICU) have heterogeneous pathophysiology, and failure to breathe without assistance from a ventilator can be the result of primary respiratory or cardiac failure, or a mixed etiology. Physicians and nurses need prediction tools to help with clinical decision making when assessing children in the CICU for extubation readiness.

This paper develops a prediction tool using large-scale "shallow" data from a clinical registry of over 32 institutions from North America (Pediatric Cardiac Critical Care Consortium: PC4), combined with small-scale "deep" data from CICU monitors and devices at 1 minute intervals. The latter data source allows the opportunity to study physiologic parameters during the key period when patients are evaluated for extubation readiness. We develop a tandem machine learning based approach to combine large-scale, shallow data with small-scale, deep data to improve prediction. The idea is to perform sequential classification: first using widely available covariates for risk stratification and subsequently refining prediction using deep data. Time series models are used to extract features from the deep data. We develop a novel framework that is time and cost-effective, for identifying patient subgroups that would most benefit from a second-stage prediction refinement using the deep data. Final tandem prediction is obtained by combining predictions from both the first and second stage classifiers. Our proposed method yields a classifier with improved prediction accuracy for predicting extubation failure in the CICU.

**AUTHORS/INSTITUTIONS:** M. Banerjee, Biostatistics, University of Michigan, Ann Arbor, Michigan, UNITED STATES|

**CONTROL ID:** 3368369

**TITLE:** Mathematical Dynamics of Parasitic Model With Migration And Induced Death Rates

**ABSTRACT BODY:**

**Abstract Body:** Malaria is an infectious disease which is caused by the Plasmodium parasite which the human population can be infected from the bites of infectious female mosquitoes. Despite the awareness, some communities are still faced with the menace of the fever caused by these ubiquitous mosquitoes.

A mathematical model describing the dynamics of the transmission to establish the potential havoc needs to be studied. In this paper, a mathematical model involving ordinary differential equations was modified to accommodate migration. The induced death rate was also incorporated into the model.

The threshold number of the model was obtained. Local and global stability was also studied.

The study revealed that migration into a susceptible population increased the threshold number. The co-infection of Malaria with the induced death was drastically influenced by the dynamics of infectious mosquitoes. Numerical simulation was used to illustrate and real-life data was also used to validate the result.

**AUTHORS/INSTITUTIONS:** M. Ekum, Maths and Statistics, Lagos state polytechni, Lagos, Lagos, NIGERIA|J.A. Akinyemi, Maths and Statistics, Lagos State Polytechnic, Ikorodu, Lagos, Lagos, NIGERIA|A. Chukwu, Statistics, University of Ibadan, Ibadan, Oyo, NIGERIA|

**CONTROL ID:** 3368378

**TITLE:** FACTORS ASSOCIATED WITH RETENTION AMONG ADOLESCENT AND YOUNG ADULTS RECEIVING ANTIRETROVIRAL THERAPY, LESSONS FROM KOGI STATE, NIGERIA

**ABSTRACT BODY:**

**Abstract Body:** Background: HIV is one of the world's most serious public health challenges causing millions of young adults' death, devastating and impoverishing families. The menace had turned millions of children into orphans. Amongst infected individual including adolescent and young adults, retention in HIV care becomes worrisome after ART initiation, which is extremely imperative not only to reduce individuals HIV-related mortality and morbidity but also as a means to deliver positive prevention intervention at reducing ongoing transmissions.

**Objectives:** The objectives of the study were to investigate factors associated with retention of HIV infected adolescents and young adults on antiretroviral treatment and their socio-demographic characteristics in the Kogi State.

**Methods:** A descriptive, cross-sectional study using a multistage sampling technique was used to select 307 respondents living with HIV and receiving antiretroviral treatment in Kogi State. **Result:** The result showed that over half (52.1%) of HIV patients were adolescents, majority (58.6%) were female and single (85.7%). Amongst these patients, approximately one-fourth (19.9%) didn't have formal education. There was significant association between lack of interest developed by these patients on ARVs and their retention in care. There was also significant association between stigmatization and patients' retention in care.

**Conclusion:** Stigmatization is a challenge to the HIV treatments, this result into patients' lackadaisical attitude and misgivings about the way HIV patients are serious about treatments.

**Keywords:** Adolescent, Young adults. Retention, Antiretroviral therapy, HIV

**AUTHORS/INSTITUTIONS:** M. Luke, Community Medicine, Ladoke Akintola University of Technology, Oshogbo, Lokoja, Kogi, NIGERIA|

**CONTROL ID:** 3368431

**TITLE:** Non-parametric confidence intervals for partial areas under the receiver operating characteristic curves and their differences

**ABSTRACT BODY:**

**Abstract Body:** Receiver operating characteristic (ROC) curves are used to describe and compare the performance of diagnostic technology and diagnostic algorithms, commonly in terms of areas under the curves (AUCs). A major drawback of an AUC is that it includes regions that may not be of relevance in practice. An alternative index is the partial AUC (pAUC), which summarizes a portion of the curve over the prespecified range of interest, e.g. a range with high specificity. Contrast to AUC, there is a lack of statistical methods for constructing confidence intervals for pAUC and differences between two pAUCs. In this work, we propose nonparametric methods for interval estimation for pAUC and differences between two pAUCs based on paired data. Due to the nonparametric nature, the methods are applicable to continuous data as well as ordinal data. Simulation results show that the methods perform very well in terms of nominal coverage and tail errors for sample sizes as small as 50. The use of the methods is illustrated in the analysis of biomarker tests for pancreatic cancer.

**AUTHORS/INSTITUTIONS:** G. Zou, Epidemiology & Biostatistics, Western University, London, Ontario, CANADA|

**CONTROL ID:** 3368434

**TITLE:** A pathway activity inference method using integrative directed random walks for predicting survival in cancer

**ABSTRACT BODY:**

**Abstract Body:** Urologic cancers include prostate, kidney, and bladder cancer with common genetic architecture in different types. To better understand the molecular features of urologic cancers, a comprehensive analysis using multi-omics data has been conducted. Additionally, a pathway activity inference method has been developed to facilitate the integrative effects of multiple genes. However, one of the limitations of the existing pathway activity inference approaches is to target a single genomic profile alone. In this respect, we have recently proposed a novel integrative analysis approach using a directed random walk-based pathway activity inference method and demonstrated that it not only contributed to a higher survival group classification performance and also successfully reflected the combined effect of genes. In this study, we integrated multi-omics data to infer pathway activities for predicting survival outcome in urologic cancers. To reflect the interaction effects of genes, we designed a directed gene-gene graph using pathway information by assigning interactions between genes in multiple layers of networks. The proposed method selects cooperative driver pathways and predicts survival outcome using Lasso-Cox model. As a proof-of-concept study, multi-omics datasets, including RNA-Seq, DNA methylation, and copy number data, for bladder cancer were obtained from the Cancer Genome Atlas (TCGA). In the experiments, the proposed integrative method on three genomic profiles was evaluated with respect to the survival outcome prediction performance and the prioritized genes and pathways for survivability of bladder cancer patients. In the results, the proposed integrative approach achieved average 7.7% improvements (C-index) of survival prediction performance than that of a single genomic profile. The integrative approach also identified 13 pathways as predictive prognostic features in bladder cancer, which are not found from a single genomic profile. They include 4 amino acid metabolism related pathways, 3 DNA repair pathways, and 6 cell proliferation, and growth-related pathways. Our results showed that the integrative approach guided by pathway information not only improves survival outcome prediction performance, but also provides better biological insights into the pathways and genes prioritized by the model in an integrated view.

**AUTHORS/INSTITUTIONS:** S. Kim, E. Choe, M. Shivakumar, D. Kim, Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES|S. Kim, K. Sohn, Department of Software and Computer Engineering, Ajou University, Suwon, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3368484

**TITLE:** Some Thoughts on the QR Method for Analytical Similarity Evaluation

**ABSTRACT BODY:**

**Abstract Body:** As indicated in a recent published draft guidance on comparative analytical assessment, the United States (US) Food and Drug Administration (FDA) seems to suggest the use of quality range (QR) method for analytical similarity evaluation. It is a concern that the use of QR method for analytical similarity evaluation could potentially approve biological products which are not deemed biosimilar to the reference biological products. In this article, the limitations and potential risk for the use of the QR method for analytical similarity evaluation are discussed. Alternatively, two modified versions of the QR method, which are referred to as effect size (ES) mQR and plausibility interval (PI) mQR methods are suggested. The performance and statistical properties of the mQR methods are evaluated via extensive clinical trial simulation under various scenarios. The results indicate that the modified versions of the QR method not only overcome the limitations of the QR method for analytical similarity evaluation, but also can potentially help in detecting reference product changes during manufacturing process.

**Key Words:** Comparative Analytical Assessment; Quality Control; the mQR Method; False Positive Rate; Reference Product Change

**AUTHORS/INSTITUTIONS:** S. Son, M. Choo, CMC Statistics, Celltrion inc., Incheon, KOREA (THE REPUBLIC OF)|M. Oh, S. Lee, Celltrion Inc., Incheon, KOREA (THE REPUBLIC OF)|S. Chow, Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, UNITED STATES|

**CONTROL ID:** 3368509

**TITLE:** Efficient Estimation of Hidden Ancestry Substructure in Summary Genotype Frequency Data

**ABSTRACT BODY:**

**Abstract Body:** Modern genomic research has advanced at a rapid pace in the last few decades, resulting in new and ever-expanding online genotype frequency databases. This publicly available, summary level data is used as controls in association studies and to prioritize possible causal variants. However, some of the data has heterogeneous ancestry, such as the African/African-American group within the Genome Aggregate Database (gnomAD). Lack of precise ancestry information can lead to confounding in association studies and incorrect prioritization of putative causal variants. Further, ancestry differences between the database and a user's sample limits the utility of the database.

We present a method to estimate the proportion of reference ancestry groups within genotype frequency databases. Our method uses sequential quadratic programming, an iterative minimization algorithm, to estimate the mixture proportions in seconds and enables us to use block bootstrapping to estimate error for the ancestry proportion estimates. We use the estimated ancestry proportions to update the database's expected allele frequency to match the ancestry of the user's sample, increasing the utility of these genetic databases for all ancestry groups.

We evaluate our method in thousands of simulation scenarios and in real data using a reference panel that includes 1000Genomes super-populations (non-Finnish European, South Asian, East Asian, and African) and Indigenous American ancestry. Across all simulation scenarios, we obtain ancestry proportion estimates with 0.05% accuracy and precision. Within the gnomAD African/African-American group we estimate ancestry percentages of 82.49% African, 15.66% European, 0.84% Indigenous American, 0.51% South Asian, 0.50% East Asian. We then use these ancestry estimates to provide updated allele frequency estimates for African ancestry. Given an accompanying R package and Shiny App, our method allows for better use of valuable genetic resources.

**AUTHORS/INSTITUTIONS:** I. Arriaga-MacKenzie, G. Matesi, A. Ronco, R. Scherenberg, A. Zerwick, J. Vance, J. Hall, Y. Wu, M.M. Null, University of Colorado Denver, Glendale, Colorado, UNITED STATES|C. Gignoux, Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Denver, Colorado, UNITED STATES|A. Hendricks, Mathematical and Statistical Sciences, University of Colorado Denver, Denver, Colorado, UNITED STATES|

**CONTROL ID:** 3368518

**TITLE:** Analysis of Gender Parity for Pakistan: Ensuring Inclusive and Equitable Quality Education

**ABSTRACT BODY:**

**Abstract Body:** Considering how much progress has been made in education in Pakistan, and how large an effort is needed to meet gender parity in primary education. Education is at the heart of sustainable improvement and the SDGs, a cause of action and hope. Educating girls as well as boys is an achievable goal and attainable in the near term if substantial resources are matched with comprehensive national strategies for education reform that include measures of accountability and a commitment to ensure every girl and boy in school. Additionally, the study signifies that how far away we are from accomplishing these SDGs. These results ought to set off alerts and prompt a noteworthy scale-up of activities to accomplish SDG 4 and ensuring gender parity. Moreover, it underlines the gaps where Pakistan stands today in education and where it has to establish reaching by 2030. Projections through regression modeling illustrate how many years will be needed to accomplish gender parity in education for Pakistan at a disaggregated level. Such a comprehension could go some approach to have the anticipated evaluations in graphics significantly lifted. While challenges still exist, the expected distance to achieve gender parity provides us guidance on to make significant progress. Punjab and urban areas have achieved gender parity for primary enrollments while other provinces need to learn lessons. An emphasis on equity is likewise be required over the full SDG motivation, as the objectives won't be achieved unless advancement is made for all least developed districts and provinces, and as a whole. In short, there may be no better investment for the health and development of Pakistan than investments to educate girls.

**AUTHORS/INSTITUTIONS:** M. Umar, Research and Survey, National Institute of Population Studies, Ministry of National Health Services, Regulations & Coordination, Islamabad, Pakistan, Islamabad, ICT, PAKISTAN|Z. Asghar, School of Economics, Quaid-i-Azam University, Islamabad, ICT, PAKISTAN|

**CONTROL ID:** 3368881

**TITLE:** Copula-based modeling of maternal antenatal exposure to ambient oxides of nitrogen with joint adverse preterm birth and low birthweight outcomes using geoadditive bivariate probit model

**ABSTRACT BODY:**

**Abstract Body:** The adverse effects of ambient air pollution on preterm birth and low birthweight are complex, and a nonlinear dependence between the covariates and the adverse birth outcomes is more realistic in addition to linear relationships. The aim of our study was to investigate the impact of maternal antenatal exposure to ambient oxides of nitrogen during pregnancy on joint adverse birth outcomes of preterm birth and low birthweight. Six hundred and fifty-six mother-newborn pairs from the Mother and Child in the Environment birth cohort from the city of Durban, South Africa participated in this study. A copula-based geoadditive bivariate probit model was used. Continuous covariates were allowed to affect the two adverse birth outcomes non-linearly using the univariate and bivariate thin plate regression splines smooth functions. Despite the geoadditive bivariate probit model being computationally intensive, a potential advantage is that the effect of air pollution can be detected with higher power under a plausible joint model rather than fitting the independence model, given the high correlation between the adverse birth outcomes and exposure to oxides of nitrogen. The results showed that of the 656 infants in the study, the observed co-occurrence of preterm birth and low birthweight was found to be 7.5%, where 17.2% were preterm and 14.2% were low birthweight. After adjusting for potential confounding factors such as socio-economic, demographic, clinical, and behavioural factors, exposure levels to oxides of nitrogen and spatial variation were found to have a non-linear effect on the joint adverse birth outcomes. In addition, maternal smoking was also found to have an increased risk of both preterm birth and low birthweight. The study suggested that the spatial identification of high risk areas for the joint occurrence of preterm birth and low birthweight gives decision makers a prompt view of air pollution in residential communities that need their attention.

**AUTHORS/INSTITUTIONS:** A.A. Mitku, T. Zewotir, D. North, Statistics, University of KwaZulu-Natal, Durban, KwaZulu-Natal, SOUTH AFRICA|A.A. Mitku, Statistics, Bahir Dar University, Bahir Dar, ETHIOPIA|R. Naidoo, Occupational and Environmental science, University of KwaZulu-Natal, Durban, SOUTH AFRICA|

**CONTROL ID:** 3369594

**TITLE:** Weighted pseudo-values for partly unobserved group membership in stem cell transplantation studies

**ABSTRACT BODY:**

**Abstract Body:** Studying paediatric leukaemia patients with and without available stem cell donors enables to compare the efficacy of stem cell transplantation with chemotherapy. Since donor search is expensive and takes time, a patient's donor availability status becomes known either when a donor is identified or when no donor is found after extensive searching. Donor availability remains unknown when donor search is cancelled for both cost and ethical reasons. This usually happens in case of premature death and when stem cell transplantation is no longer considered a suitable treatment option (e.g. after a relapse). Technically, donor availability status can be considered as a partly unobserved, external, binary time-dependent covariate. The statistical situation becomes even more challenging due to non-proportional hazards resulting from an increased post-transplantation mortality that vanishes over time. The generalised pseudo-values approach is able to correctly address these issues and enable an unbiased treatment comparison. However, calculations are time-consuming due to a necessary bootstrap step. A faster alternative approach is suggested which is based on weighting of common pseudo-values and does not require resampling. Both approaches are compared by a simulation study where they show similarly satisfactory behaviour with respect to confidence interval coverage and type I errors. A real data example is presented and analysed to illustrate and discuss both approaches.

**AUTHORS/INSTITUTIONS:** U. Pötschger, CCRI, Vienna, AUSTRIA|M. Mittlböck, H. Heinzl, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, AUSTRIA|

**CONTROL ID:** 3369917

**TITLE:** Dimension Reduction using Local Principal Components in Multi-SNP Genetic Association Analysis

**ABSTRACT BODY:**

**Abstract Body:** Combining genetic association effect estimates of multiple single nucleotide polymorphisms (SNPs) within a genomic region is an effective strategy to deal with high-dimensional genotype data with complex correlation structure since it can augment the strength of signals while reducing the dimension of the analysis. One such method first takes the results of multi-SNP regression analysis and then constructs single or multiple degree-of-freedom (df) Wald tests including single linear combination (SLC) tests and multiple linear combination (MLC) tests that combine the resulting regression coefficients. Alternatively, regional statistics constructed from a principal component regression (PCR) can also be seen as a multiple-regression-based approach with reduced df which first transforms the original SNP variables using linear combinations before assessing association with the phenotype of interest. PCR can resolve the multi-collinearity caused by regression of many SNP variables with high correlation. However, multiple principal components are hard to interpret as a biological entity. In this study, we propose an algorithm DRLPC (Dimension Reduction using Local Principal Components) to reduce the dimension for regional regression analysis by identifying clusters of SNPs with high correlation and replacing each cluster by the first local PC constructed from the SNPs in the cluster, which improves interpretability. Any multicollinearity remaining among these updated cluster-specific variables is resolved by considering variance inflation factors (VIF) and removing variables with high VIF. We investigated the performance of DRLPC using the local PCs in multi-SNP tests when applied to the genotype data prior to multiple regression. Simulation studies based on chromosome 22 data of 1000 Genomes Project revealed that Wald test power using DRLPC is similar to the MLC test power when the same threshold values are used for clustering. That is, the order of applying clustering and combining SNPs has modest effects on hypothesis testing. We conclude that DRLPC can provide efficient dimension reduction while resolving complex multi-collinearity and also improves interpretability because these principal components represent subsets of SNPs in the region, possibly short haplotypes.

**AUTHORS/INSTITUTIONS:** Y. Yoo, F. Yavartanoo, Mathematics Education, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|Y. Yoo, 2Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|S.B. Bull, Lunenfeld-Tanenbaum Research Institute, University of Toronto, Toronto, Ontario, CANADA|S.B. Bull, 4Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, CANADA|

**CONTROL ID:** 3370527

**TITLE:** STATISTICAL ANALYSIS OF PALMAR AND DIGITAL DERMATOGLYPHIC TRAIT ASYMMETRY AMONG TYPE 2 DIABETIC & NON DIABETIC ADULTS: A META ANALYSIS

**ABSTRACT BODY:**

**Abstract Body:** BACKGROUND

Dermatoglyphic patterns are the epidermal ridges seen on the surface of palm, sole & digits. These ridges play a significant role in assessing various diseases in mankind. Diabetes in today's world is a challenge and serious threat as lifestyle disorder. It is very important to know about the early diagnosis and undertake the preventive measures to overcome the threat. The aim of this review is to establish statistical significance with various dermatoglyphic studies and find out the significance results in the literature which shows the association between dermatoglyphic & diabetes mellitus. Additionally, this review aims to systematically identify, review and appraise available literature that evaluate an association of different dermatoglyphic variables with kidney diseases.

**METHODS**

An intense systematic literature search was conducted using keywords 'Dermatoglyphic' and diabetes from Medline (PUBMED), Google Scholar, EBSCO, HINARI etc. The review is performed based on the Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) statement. Dermatoglyphic patterns atd, dat, adt angles, absolute finger ridge count (AFRC), total finger ridge count (TFRC), ab ridge count, mainline index, and pattern index line were studied.

**RESULTS**

The mean values of TFRC and AFRC was higher in male and lower in the female diabetic group. The mean values of ab ridge count was lower in male and higher in the female diabetic group and a statistical significant difference was found. The mean values of atd and adt were higher in the diabetic group.

**CONCLUSION**

This review and metaanalysis shows statistically stronger result due to increased numbers of cases, greater diversity among subjects, confirmatory data analysis and pooled results. This study alliance can be explained and justified if the risk towards developing Diabetes mellitus in future life could be connected with the fetal development of epidermal ridges. Dermatoglyphics provide a simple, inexpensive, anatomical, and noninvasive means of determining the diseases which have a strong hereditary basis and can be employed as a method of screening for diabetes mellitus of the high risk population on early detection, thus reducing the morbidity and mortality.

**AUTHORS/INSTITUTIONS:** A.K. Jha, Preclinical Sciences, Texila American University, East Bank Demerara, GUYANA|S. D'costa, American University of Antigua, Coolidge, ANTIGUA AND BARBUDA|

**CONTROL ID:** 3370715

**TITLE:** Multivariate Analysis of Correlated Mass Spectrometry data with Non-Random Missing Values

**ABSTRACT BODY:**

**Abstract Body:** Combining information from multiple experiments that address the same or correlated hypothesis can provide meaningful answers to important biological questions. We restrict our attention to situations in which the multiple experiments aim to determine common genomic variants across multiple tissue types. For example, our cancer biomarker project entails profiling several biological specimens (e.g., tissue, plasma, serum, and urine) from the same subjects. Mass spectrometry (MS) is now being used to profile small molecular compounds across multiple biological sample types. Multivariate statistical methods that combine information from all biological samples could be powerful than the usual univariate analyses separately. Missing values are common in MS data and imputation can impact between samples within subject correlation and thereby multivariate analysis results. Several underlying mechanisms have been identified in driving this missingness: a compound may either be present in a sample but at a concentration below the limit of detection (LOD) of the mass spectrometer, completely absent from a sample, or be above the LOD but not detected due to technical limitations. Thus, missingness in MS data is a combination of "missing completely at random (MCAR)" and "missing not at random (MNAR)," meaning many traditional imputation methods are not applicable due to their reliance on random missingness. In most cases, peak intensities are not measured because their abundance is simply below the detection limit of the mass spectrometer. Herein, we propose multi-biospecimen testing procedures using two part statistics that accommodate non-random missing values and combine from all biological samples within subject to identify differentially regulated compounds. Simulation studies are carried out to test the performance of the proposed approaches and compare with other existing multivariate analysis methods.

**AUTHORS/INSTITUTIONS:** K. Kim, S.L. Taylor, Biostatistics, University of California- Davis, Davis, California, UNITED STATES|

**CONTROL ID:** 3372121

**TITLE:** ROBUST ESTIMATION OF SINGLE-INDEX MODELS WITH RESPONSES MISSING AT RANDOM

**ABSTRACT BODY:**

**Abstract Body:** A single-index regression model is considered, where some responses in the model are assumed to be missing at random. Local linear rank-based estimators of the single-index direction and the unknown link function are proposed. Asymptotic properties of the estimators are established under mild regularity conditions. Monte Carlo simulation experiments show that the proposed estimators are more efficient than their least squares counterparts especially when the data are derived from contaminated or heavy-tailed model error distributions. When the errors follow a normal distribution, the least squares index direction estimator tends to be more efficient than the rank-based index direction estimator; however, the least squares link function estimator remains less efficient than the rank-based link function estimator. A real data example is analyzed and cross-validation studies show that the proposed procedure provides better prediction than the least squares method when the responses contain outliers and are missing at random.

**AUTHORS/INSTITUTIONS:** M. Otladisa, B. Makubate, Department of Mathematics and Statistical Sciences, Botswana International University of Science & Technology (BIUST), Palapye, BOTSWANA|A. Abebe, Department of Mathematics & Statistics, Auburn University, Auburn, Alabama, UNITED STATES|H.F. Bindele, Department of Mathematics & Statistics, University of South Alabama, Mobile, Alabama, UNITED STATES|

**CONTROL ID:** 3372144

**TITLE:** Causal Inference in discrete latent variable model.

**ABSTRACT BODY:**

**Abstract Body:** This paper discusses the causal inference between two discrete latent variables. Suppose that a researcher is studying for the causal effect of substance use behavior on the violent behavior of the high school students. In this question, both substance use and violent behavior are categorical latent variables, because they cannot be observed directly but can only be measured via a set of questionnaire variables. To investigate the pure causal effect of treatment on outcomes, the effect of other existing covariates on the causal relationship between treatment and outcome variables should be adjusted, because the compounding effect of the covariates may distort the pure effect of treatment on outcome. Propensity score had been used to address the covariate compounding effect issues. In this sense, we suggest two strategies for the causal inference in latent class models. One is the Latent Class Causal Model (LCCM) which provides the causal relationships between two unobservable categorical latent variables, where each latent variable is identified from discrete response variables. LCCM suggests a new statistical approach to the causal inference in discrete latent variables using the propensity score. It can be divided into two steps. In the first step, the propensity scores can be obtained multinomial logistic regression by regressing the covariates on the treatment latent variable. Here, the treatment latent variable can be modeled based on latent class analysis or joint latent class analysis (Jeon, 2017). The second approach is propensity score matching to balance the compounding effect of covariates on the causal relations. Once the propensity scores are obtained from step 1, we divide the sample into several subgroups so that the observations within the same groups have a similar propensity score. Then, within each cluster, we fitted the LCAG model and combined the estimates from each subgroup. We implemented a numerical simulation whether two strategies properly provide the parameter estimates. We found that LCCM with inverse propensity score weight provides stable estimates, while the estimates via propensity score matching method are unstable if the number of groups is insufficient to balance the covariate effects. For the real data example, we investigated the causal influence of substance use behavior on the violence behavior among US high school students.

**AUTHORS/INSTITUTIONS:** J. Lee, Statistics, Univerisity of Connecticut, Willington, Connecticut, UNITED STATES|

**CONTROL ID:** 3372407

**TITLE:** A Generalized Linear Model for Estimating Diabetes and Hypertension Care Costs

**ABSTRACT BODY:**

**Abstract Body:** The prevalence of diabetes in Nigeria adults is two per cent with total cases of 1,702,900. More so, hypertension and diabetes mellitus also account for 7% and 2% of the 2.08 million deaths attributable to Non Communicable Diseases in Nigeria. To maintain quality life after being diagnosed, patient would continually incur healthcare costs. This study develops Generalized Linear Models for estimating costs of maintaining patients with diabetes and hypertensive conditions. The healthcare utilization and costs data used in this research are characterized by highly non-Gaussian distribution with severe skewness and kurtosis. We examined a variety of models commonly used to determine the most appropriate model specification within the Poisson, negative binomial, or gamma variance functions.

**AUTHORS/INSTITUTIONS:** L.A. Ajijola, Actuarial Science and Insurance, University of Lagos, Lagos, NIGERIA|I.A. ADELEKE, Actuarial Science & Insurance, University of Lagos, Lagos, NIGERIA|

**CONTROL ID:** 3372673

**TITLE:** Study of allergic rhinoconjunctivitis in childhood through fuzzy clustering methods for functional data

**ABSTRACT BODY:**

**Abstract Body:** Allergic Rhinoconjunctivitis (AR) is a common disease in children. Although often undiagnosed, AR affects the quality of life because it may cause several comorbidities including asthma. There exist several indexes for monitoring AR. They measure the levels of the perceived severity of various symptoms and the possible consumption of anti-symptomatic drugs. In this work we propose the use of fuzzy clustering methods for functional data in order to study the daily values of the so-called Combined Symptom and Medication Scores (CSMS) index observed on two populations of patients with seasonal allergic rhinitis. Under the assumption that such measurements arise from smooth continuous functions, functional data are generated and then clustered in order to discover patients with similar dynamic evolutions of CSMS. The obtained groups may be useful for diagnostic classification and cluster-specific therapies.

**AUTHORS/INSTITUTIONS:** P. Giordani, Sapienza University of Rome, Rome, ITALY|S. Perna, P. Matricardi, Charitè Medical University of Berlin, Berlin, GERMANY|A. Bianchi, San Camillo Forlanini, Rome, ITALY|A. Pizzulli, Practice of Pediatric Pneumology and Allergology, Berlin, GERMANY|S. Tripodi, Sandro Pertini Hospital, Rome, ITALY|

**CONTROL ID:** 3373051

**TITLE:** A Mixed-Integer Linear Programming Approach to Two-dimensional Arrays for Two-Phase Experiments

**ABSTRACT BODY:**

**Abstract Body:** Design of experiments is widely used in many areas (e.g. industrial engineering, agriculture, medical research, etc.). Often, experiments involve two different phases. For example, in plant breeding the first phase (Phase 1) of the experiment may be performed in a field involving a number of treatments e.g. varieties) and a single blocking factor (i.e. field blocks), whereas the second phase (Phase 2) may be performed in a lab to measure the response using the samples from the first phase, taking into account the presence of another blocking factor (e.g. days or lab machines). Such experiments are referred as to two-phase experiments. In this talk, we focus on a class of designs for two-phase experiments that involve one treatment factor and a single blocking factor in each phase. Additionally, we assume that the same number of experimental units is used in both phases, and that all treatments have the same replication. We say that such designs belong to class C. In this talk, we focus on threefold. First, under some special design settings, we illustrate that the two-phase designs in class C can be constructed from triple-arrays (Wallis, 2014) and sesqui-arrays (Bailey, Cameron, and Nilson, 2018). Second, we propose a mixed-integer linear programming approach (MILP) to construct such arrays. Finally, we demonstrate the capacity of our MILP approach with various examples, and end with a discussion.

**References**

Bailey, R.A, Cameron P.J, and Nilson T (2018). Sesqui-arrays, a generalisation of triple arrays. *Australasian Journal of Combinatorics*, 71: 427-451.

Wallis, W. D. (2014). Triple arrays and related designs. *Discrete Applied Mathematics*, 163: 220-236.

**AUTHORS/INSTITUTIONS:** H. Piepho, N. Vo-Thanh, Biostatistics Unit, Institute of Crop Science, Stuttgart, GERMANY|

**CONTROL ID:** 3373363

**TITLE:** Identification of Effect Modifier in Multidrug-Resistant Tuberculosis in an Individual Patient Data Network Meta-Analysis

**ABSTRACT BODY:**

**Abstract Body:** Multidrug-resistant tuberculosis (MDR-TB) is caused by an infection which is resistant to at least the two most effective anti-TB drugs, isoniazid and rifampin. The treatment of MDR-TB is more challenging than drug sensitive TB because different patients may be resistant to various drugs and are typically prescribed a combination of four or more antimicrobial agents. Previous work investigated average treatment effectiveness mainly based on small sample observational studies and subsequent individual patient data (IPD) meta-analyses of those studies (Ahuja et al. 2012; Ahmad et al. 2018). Effect modification occurs while the effect of the treatment on an outcome is not homogeneous in the different strata formed by patient characteristics. If effect modification presents, estimating an overall treatment effect might provide less guidance for individual treatment since it may not characterize the effect of the treatment for any given patient. The common way of dealing with effect modification is including interaction terms between the treatment and the potential effect modifiers in a regression model and proceeding with strata-specific estimation. However, this method relies on a correct model for the outcome.

The use of marginal structural models (MSMs) to adjust for measured confounding factors is becoming more common in observational studies. In this paper, we propose MSMs to identify the effect modifiers in a network meta-analysis of observational IPD. Our data arise from a systematic review by Ahuja et al. (2012) and our analysis was undertaken using IPD of MDR-TB patients from 31 studies. We develop a novel doubly robust implementation of Targeted Maximum Likelihood Estimation (TMLE) to assess effect modification by different patient characteristics and co-medications in an MSM in the IPD meta-analysis. TMLE is a loss-based semi-parametric estimation method that yields a consistent estimator of the parameter of interest under regularity conditions by solving the efficient influence curve estimating equation. Moreover, in this analysis, we allow for differential availability of treatments across studies and random effects by study due to measured and unmeasured characteristics of the study-specific populations.

**AUTHORS/INSTITUTIONS:** Y. Liu, A. Benedetti, Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, CANADA|G. Wang, Epidemiology, Biostatistics and Occupation Health, McGill University, Montreal, Quebec, CANADA|M. Schnitzer, Faculty of Pharmacy, Université de Montréal, Montreal, Quebec, CANADA|M. Schnitzer, Department of Social and Preventive Medicine, Université de Montréal, Montreal, Quebec, CANADA|E. Kennedy, Statistics Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, UNITED STATES|D. Menzies, McGill University Health Centre, Montreal, Quebec, CANADA|D. Menzies, Department of Medicine, Mcgill University, Montreal, Quebec, CANADA|

**CONTROL ID:** 3373771

**TITLE:** Bi-serial rank correlation analysis of case findings for strategic HIV Testing approaches in some selected Health facilities, Kogi State, Nigeria

**ABSTRACT BODY:**

**Abstract Body:** Background: The 2019 NAHS indicates that Kogi State has a HIV prevalence of 0.9% with a HIV burden of 40,288 population of people living with HIV. However, only 23,768 No. population of PLHIV on treatment representing 59% with a gap of 16,520 No PLHIV yet to be diagnosed and linked-to-care. The first of the global 90:90:90 targets demands that HTS coverage must increase within the geographic areas and populations where the HIV burden is highest, and in previously under-served areas and populations at risk for HIV.

Targeted and Innovative HIV testing approaches are required alongside robust M&E systems that evolve to provide the information to enhance and focus testing efforts in order to meet this challenge.

**Methods:** The study was conducted through a rank bi-serial comparison of the correlation coefficient derived from the comparison of the yield of case findings for 2018 and 2019 HTS at 6 selected HCFs. To ensure quality, the data was validated prior to the comparison of the rank correlation coefficient analysis for 2018 and 2019. Targeted PITC, Integrated service delivery (one-stop model), Home-based testing, Health fairs and multi-disease screening campaigns/events, Index-client testing and partner notification services provided the approach for the identification of case findings. Data was collected on client intake forms and HTS register for the routine diagnostic tests for HIV 1 and 2 for clients by HCFs.

**Results:** A trend lines  $Y = 21.1 + (0.002) X$  and  $Y = 51.6 + (0.001) X$  were derived. The analysis indicates that the RDTs for HIV outcomes for some selected health care facility presents  $r^2$  values of 0.53 and 0.10 for 2018 and 2019 respectively for PLHIV has had a 43% drop in the usage of RTKs while increasing the yield of case findings.

**Conclusions:** The inference has an importance that has shifted the relevance of HTS towards greater efficiency due to the targeted strategic drive. 1 No. yield in 2018 was dependent on 53 RDTs while a 1 No. yield in 2019 was dependent on 10 RDTs, thus these findings emphasize targeted RDTs for improving the yield by 43% current assets utilization (RTKs) reduction.

**Keywords:** Yield; Targeted HTS; case-finding, Rank Correlation Coefficient

**AUTHORS/INSTITUTIONS:** M. Luke, Community Medicine, Ladoke Akintola University of Technology, Oshogbo, Oshogbo, Osun State, NIGERIA|M. Luke, Monitoring and Evaluation, AIDS Healthcare Foundation, Lokoja, Kogi, NIGERIA|

**CONTROL ID:** 3374025

**TITLE:** Bayesian Decision Fusion of Palm and Finger Prints for Personal Identification

**ABSTRACT BODY:**

**Abstract Body:** The ever increasing demand of security has resulted in a wide use of Biometrics systems. Despite overcoming the traditional verification problems, the unimodal systems suffer from various challenges like intra class variation, noise in the sensor data etc, affecting the system performance. These problems are effectively handled by the multimodal systems. In this paper, we present a multimodal approach for palm and finger prints by decision level fusion. The proposed multi-modal system is tested on a developed database consisting of 440 palm and finger prints each of 55 individuals. In methodology of decision level fusion, Directional energy based feature vectors of palm and finger print identifiers are compared with their respective databases to generate scores based upon which final decisions are made by the individual matcher which are combined through Bayesian decision fusion. Receiver Operating Characteristics curves are formed for the unimodal and multimodal systems. Equal Error Rate (EER) of 0.7321% for decision level fusion shows that the multimodal systems significantly outperforms unimodal palm and finger prints identifiers with EER of 2.822% and 2.553% respectively.

**AUTHORS/INSTITUTIONS:** J. Hashmi, M. Mumtaz, National University of Sciences and Technology, Islamabad, Pakistan, Rawalpindi, PAKISTAN|

**CONTROL ID:** 3374372

**TITLE:** Filtering texture effects in dyed textiles images using hidden Markov random fields.

**ABSTRACT BODY:**

**Abstract Body:** Color is one of the most important features in any textile material. Due to its competitive price, most of the colorants currently used for textile dyeing are synthetic, originated from non-renewable sources and highly pollutant. There is an increasing interest for natural processes to dye fabrics. When new textile dyeing technologies are developed, evaluating the quality of these techniques involves measuring the resulting color homogeneity using digital images. The presence of a texture effect, caused by the interlacing of warp and weft yarns as well as small displacement of the fabric, creates a sophisticated dependence structure in pixels coloring. A random effects model is employed in order to separate the signal from the dyeing effect (fixed effect described by smooth functions) and warp and weft texture effect (Gaussian mixture driven by a hidden Markov random field), allowing an evaluation of color homogeneity in dyed textiles regardless of the effect of the texture.

**AUTHORS/INSTITUTIONS:** N.L. Garcia, V. Freguglia, Statistics, University of Campinas, Campinas, Sao Paulo, BRAZIL|J. Bicas, Food Sciences, University of Campinas, Campinas, BRAZIL|

**CONTROL ID:** 3374416

**TITLE:** The Impact of Design Misspecifications on Survival Endpoints

**ABSTRACT BODY:**

**Abstract Body:** Overall survival is an important endpoint in confirmatory Phase III clinical trials in oncology. Treatment effects on survival can be evaluated by comparing the mortality hazards using the log-rank test or by comparing the restricted mean survival times, or by comparing the survival probabilities at a fixed time point between the treatment arms. Given that the trials are designed based on a set of assumptions about the survival pattern of the control arm and the effect of the new treatment on survival, the trial conclusions may be erroneous when these assumptions are different from the true underlying survival patterns. The objective of this work is to evaluate the impact of misspecified design assumptions on trial conclusions. Specifically, we compare the effect of design misspecification on the statistical power using the log-rank test, the restricted mean survival time test, and the survival probability at a fixed time test.

Results of our simulations showed that:

1. The log-rank test provides the most statistical power except in scenarios where there is an early treatment effect. The restricted mean survival time test out performs the log-rank test in this setting and is the most robust against this deviation from the proportional hazards assumption. Whereas, the fixed-time survival probability test is the most robust against late treatment effect.
2. The log-rank test is the most robust against deviations from baseline hazard rate.
3. The restricted mean survival time test is the most robust against deviations in accrual rate and total trial duration.
4. Finally, the size of the treatment effect has similar impact on all three tests.

The performance of these endpoints is especially relevant in the current era of immuno-oncology and cellular therapy where the new treatment can be effective in only a subset of patients and resulting in durable response. Furthermore, there is typically a delayed efficacy and safety effect with these novel treatments making traditional approaches such as the log rank test no longer the most appropriate endpoint. Knowledge of the impact of design misspecifications on trial conclusions allow investigators to choose the most robust endpoint for their trial.

**AUTHORS/INSTITUTIONS:** J. Le-Rademacher, T. Zemla, Q. Duong, S.J. Mandrekar, Health Sciences Research, Mayo Clinic, Rochester, Minnesota, UNITED STATES|

**CONTROL ID:** 3374891

**TITLE:** VARIANCE ESTIMATION FOR METHODS OF SAMPLE PAIRING IN DUAL CHANNEL MICROARRAY APPLIED ON SPLIT-PLOT DESIGN

**ABSTRACT BODY:**

**Abstract Body:** A well-designed experiment is an important and fundamental step needed in planning a dual channel microarray experiment used in measuring gene expression. Dual channel microarray applied on split-plot design is complicated, due to the fact that important effects may be confounding in the array during the process of pairing the samples. Forming an appropriate model and using sufficient replication for estimating effects confounded with array overcomes this challenge. In this work, three methods of sample pairing which we called vertical loop method (design A), cross loop method (design B) and horizontal loop method (design C) were used in pairing samples in a dual channel microarray performed in a split-plot design in order to ascertain which method gives the minimal variance for the effects of interest. Also, the numbers of replication were varied in order to check its effect on the estimated variance. The brute force and analysis of variance methods were used in estimating the variance estimates and components in each of the designs. The results showed that design A had the least variance for comparing the mean difference between the sub-plot-treatment, colour, first and second levels of the sub-plot treatment at each level of the whole-plot treatment, and first and second levels of the whole-plot treatment at each level of the sub-plot treatment while design C had the least variance for comparing the whole-plot treatment and the variance estimated decreases as the number of replication increases. We, therefore conclude based on the results that the vertical loop method of sample pairing gave the least minimal variance when comparing the effects of treatment comparison.

**AUTHORS/INSTITUTIONS:** A. Oladugba, I. Ude, University of Nigeria, Nsukka, Nsukka, NIGERIA|

**CONTROL ID:** 3374923

**TITLE:** Genetic epistasis analysis based on the spatial rank statistic for multiple quantitative trait

authors list : Hoe-Bin Jeong, Jong-Hyun Lee, Taesung Park and Mira Park

**ABSTRACT BODY:**

**Abstract Body:** Complex diseases, such as hypertension, autism, and chronic kidney disease (CKD) are not defined by a single locus or an environmental factor. One way to find the missing heritability is to detect the genetic interactions, which is also known as genetic epistasis. The multifactor dimensionality reduction (MDR) is a well-known non-parametric approach to detect gene-gene interactions (GGIs) for binary trait. The MDR reduces the dimensions by converting a high-dimensional model to one-dimensional. Various genome-wide association studies (GWAS) still focus on a single trait to identify the genetic variants despite the availability of multiple phenotypes. However, the potential for detection of the associations between genes and diseases could be increased by modeling multivariate disease-related traits.

In the present study, we proposed a novel method for detecting GGIs for multivariate traits. We combined fuzzy-clustering and spatial rank statistic for the measurements. The rank-based statistic can get robustness from outlier and normality assumption. Two simulation studies were conducted to compare the performances of the proposed multivariate rank-based multifactor dimensionality reduction (MR-MDR) with existing methods. The performance of MR-MDR showed that it can contribute in revealing the missing heritability. We have illustrated the proposed MR-MDR method by analyzing Korean genome and epidemiology study (KoGES) data containing multiple phenotypes related to kidney function. The R program to perform the proposed MR-MDR is available.

**AUTHORS/INSTITUTIONS:** H. Jeong, J. Lee, Statistics, Korea university, Seoul, KOREA (THE REPUBLIC OF)|M. Park, Department of Preventive Medicine, Eulji university, DaeJeon, KOREA (THE REPUBLIC OF)|T. Park, statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3374924

**TITLE:** Optimal designs and efficiency of designs for state equations in distillation experiments

**ABSTRACT BODY:**

**Abstract Body:** In the distillation processes it is very important to know precisely the relationship between temperature and vapor pressure. The vapor pressures not only depend on the temperature, but vary enormously for different substances.

The Antoine equation is a hyperbolic equation, from a class of semi-empirical correlations that describe the relationship between temperature and vapor pressure very precisely for certain temperature ranges. It has two parameter sets, a low-pressure set up to the boiling point, and a high-pressure one from the boiling point onwards. The DIPPR 101 is another empirical equation, developed by the American Institute of Chemical Engineers. It is used for describing the relationship between the same physical properties, with a single parameter set of higher dimensions.

In this work, the study of optimal designs for the estimation of the parameters of both equations, according to different relevant criteria, is presented. An analytical solution for the D-Optimal designs for Antoine's Equation is shown, as well as several numerically calculated optimal designs for this and other criteria as I- and Ds- optimality. Of special interest for Antoine's Equation is the work done in I-Optimality, due to the importance of accurate predictions on boundary regions of the design space, which usually correspond to regions close to the change of state points.

The work also features an efficiency comparison between the optimal design and some industry-oriented designs with fixed number of points, and a preliminary approach of using splines to calculate optimal designs for both parameter sets of Antoine's equation together.

**AUTHORS/INSTITUTIONS:** C. de la Calle-Arroyo, L. Rodríguez-Aragón, Mathematics, Universidad de Castilla-La Mancha, Toledo, Castilla-La Mancha, SPAIN|J. López-Fidalgo, Universidad de Navarra, Pamplona, Navarra, SPAIN|

**CONTROL ID:** 3375331

**TITLE:** Joint Frailty Modeling of Recurrent and Terminal Events for Cancer Survival Risk Estimation

**ABSTRACT BODY:**

**Abstract Body:** The risks for recurrence and time to event (death) were usually estimated using Cox-proportional hazards (PH) models independently. However, terminal event is significantly linked with recurrence status. Apart from this correlated dependence between these events, there exist hidden random factors which are heterogeneous in nature. Hence to get a reliable and efficient risk estimates, in the present study the recurrent and terminal events were model jointly with random unobservable factor, the frailty. The developed joint frailty model was illustrated using breast cancer data. The frailty parameter was estimated to be 0.492, which indicates the heterogeneity between subjects explained by non-observed covariates and the positive value of the coefficient alpha in the joint model indicates that the incidence of recurrences is positively associated with death (SE: 0.192,  $p = 0.005$ ; alpha: 0.783, SE: 0.167,  $p = 0.0001$ ). The HR for stage II compared with stage I is 1.56 with 95% CI (0.81, 2.99), for stage III compared with stage I is 2.06 (1.02, 4.17) and for stage IV HR is 6.41 (2.55, 16.1). Thus the present study assessed the role of joint frailty modeling of breast cancer patient by take into account the recurrence and terminal events together. The joint models were analyzed using R program. On comparison of the joint frailty models with Cox and frailty models revealed that, for correlated time to event data with hidden heterogeneity, joint frailty model is the apt choice.

**AUTHORS/INSTITUTIONS:** J. KM, A. Mathew, P. Sara George, Division of Cancer Epidemiology & Biostatistics, Regional Cancer Centre, Thiruvananthapuram, Kerala, India, Thiruvananthapuram, Kerala, INDIA|A. Maria Anto, Biostatistics, St. Tomas College, Kottayam, Kerala, INDIA|

**CONTROL ID:** 3376114

**TITLE:** A solution to separation or near-to-separation in the multinomial logit models with application to childhood health seeking behavior data

**ABSTRACT BODY:**

**Abstract Body:** Logistic regression models are commonly applied to analyze multinomial data frequently arising in many areas of research including medicine and social sciences. However, separation leading to monotone-likelihood created by a covariate often exists in multinomial logistic model when the sample size is relatively small, the certain response categories is rare, and if there is sufficient number of influential covariates. In the presence of separation, the maximum likelihood (MLE) may fail to achieve convergence or provide biased or infinite estimate of at least one regression coefficients of the model, particularly for the coefficient associated with the covariate responsible for creating separation. This study addressed the problems of separation by applying a penalized likelihood (PMLE) approach, which was originally proposed by Kosmidis and Firth (2011) to remove the first-order bias in the MLEs of the multinomial logit model via Poisson log-linear model. The penalized likelihood is shown to achieve convergence and provide finite estimate of the regression coefficient in the presence of separation. We investigated the performance of both MLE and PMLE using an extensive simulation study against scenarios with complete (e.g., more than one empty cell) or quasi-complete-separation (one empty cell) and a new form termed as “near-to-separation” (non-zero cell but with few observations-less than 15% of total sample), which is more common in practice than the other two forms. The simulation study showed that the MLE failed to achieve convergence and/or provided infinitely large estimate of the regression-coefficient in the presence of complete or quasi-complete-separation, whereas the PMLE showed some improvements by achieving convergence and providing finite estimate. In the presence of near-to-separation, the PMLE also outperform the MLE by providing smaller amount of bias and MSE and better coverage. An application of the method is provided to analyze childhood health seeking behavior data consisting of different forms of separation.

Keywords: monotone likelihood, Poisson log-linear model, penalty function.

**AUTHORS/INSTITUTIONS:** N. Nusrat, M.S. Rahman, Institute of Statistical Research and Training, University of Dhaka, Dhaka, BANGLADESH|

**CONTROL ID:** 3376559

**TITLE:** Approaches for Extending Multiple Imputation to Handle Scalar and Functional Data

**ABSTRACT BODY:**

**Abstract Body:** Missing data are a common problem in biomedical research. Valid approaches for addressing this problem have been proposed and are regularly implemented in applications where the data are exclusively scalar-valued. However, with advances in technology and data storage, biomedical studies are beginning to collect both scalar and functional data, both of which may be subject to missingness. We propose extensions of multiple imputation with predictive mean matching and imputation by local residual draws as two approaches for handling missing scalar and functional data. The two methods are compared via a simulation study and applied to data from a study of subjects with major depressive disorder for which both clinical (scalar) and imaging (functional) data are available.

**AUTHORS/INSTITUTIONS:** A. Ciarleglio, Biostatistics and Bioinformatics, George Washington University, Washington, District of Columbia, UNITED STATES|

**CONTROL ID:** 3377163

**TITLE:** Profiles of lifestyle and metabolic risk factors for transitions of pre-diabetes status using hidden Markov models

**ABSTRACT BODY:**

**Abstract Body:** Dietary intake and physical activity that consist lifestyle profile may be modified with time to improve the status of obesity and cardiovascular risk factors and associated glucose status. Motivated by a population-based longitudinal health database, we identified different states of lifestyle profile, obesity, and profiles of cardiovascular risk factors by using hidden Markov model. We then delineated the lifestyle profile associated with changes in states of obesity and cardiovascular profile and their associations with the transitions of glucose status by using mixed-effect regression model.

**AUTHORS/INSTITUTIONS:** C. Chen, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, TAIWAN|

**CONTROL ID:** 3378111

**TITLE:** Challenges in handling missing data in untargeted LC-MS metabolomics experiments

**ABSTRACT BODY:**

**Abstract Body:** Untargeted liquid chromatography-mass spectrometry (LC-MS) metabolomics allows relative quantitation of the entire metabolome; however, such experiments pose methodological and statistical challenges. The resulting data have a high proportion of missing values, caused by biological mechanisms, technical issues, concentrations below the detection limit, and errors converting the MS signal to a numeric representation. Because the amount of missing data is large, and the fact that some metabolites are only found in one study group, excluding metabolites with missing data limits the information gained from the untargeted approach. We compared several common imputation methods (half minimum, mean, K-nearest neighbors, random forest, and quantile regression imputation of left-censored data) in analysis of untargeted LC-MS blood metabolomics in 16 adolescent girls with polycystic ovary syndrome and 5 without. We compared the results to a Bernoulli/lognormal mixture model which explicitly models the probability of the presence of the metabolite.<sup>1</sup> There were 6,747 measured metabolites; 4256 metabolites had complete data. 247 metabolites were excluded from analysis using the 80% rule; of the remaining metabolites, 254 had more than 50% missing values. Metabolite intensity was negatively correlated with the number of missing values ( $r=-0.06$ ,  $p<0.001$ ). We compared the groups using t-tests and the Benjamini-Hochberg procedure to control for multiple testing. The imputation methods resulted in between 167 (mean) and 210 (K-nearest neighbors) significantly different metabolites, compared to 271 by the mixture method approach. Analysis of untargeted LC-MS data is sensitive to the methods used to handle missing values, and consideration of these issues is an important step in the data processing and analysis pipeline.

<sup>1</sup>Nodzinski M, et al. Metabomxtr: an R package for mixture model analysis of non-targeted metabolomics data. *Bioinformatics* 30(22): 3287-3288.

**AUTHORS/INSTITUTIONS:** L. Pyle, A. Carreau, H. Rahat, G. Yesenia, K. Nadeau, M. Cree-Green, Pediatrics, University of Colorado School of Medicine, Aurora, Colorado, UNITED STATES|L. Pyle, Biostatistics and Informatics, Colorado School of Public Health, Aurora, Colorado, UNITED STATES|K. Nadeau, M. Cree-Green, Center for Women's Health Research, Aurora, Colorado, UNITED STATES|

**CONTROL ID:** 3378235

**TITLE:** An Augmented Reality Based Service Design Model for Digital Fabrication

**ABSTRACT BODY:**

**Abstract Body:** Digital fabrication is widely used in many areas because of its flexibility in production is suitable for contemporary emphasis on small-volume, large-variety production and customization. Following the booming development of digital fabrication tools, users can self-fabricate their own unique personalized products. However, existing digital fabrication systems are far more professional and complicated for novice users to operate properly, which leads to the less popularity of digital fabrication processes. In contrast, augmented reality can provide information not directly available in the real world by visualized, instructional and navigational functions to improve the interaction between humans and real objects.

This paper explored the application of augmented reality to digital fabrication and evaluated user experience in information transferring and integration by augmented reality. Users can create and improve designs by clicks-and-mortar technological training with augmented reality that lowering the threshold of technological adoption. Augmented reality can be used to learn basic skills of digital fabrication and to access necessary tools for the fabrication process. User experience can thus be easily passed on and the cohesion of different social groups can be increased. The purposes were to make digital fabrication more popular and more people are willing to turn their ideas into physical prototypes for market testing. A new type of business could therefore be envisioned.

The focus of this paper was on the establishment of a service design model for digital fabrication. Service design processes and tools were used to identify problems of current digital fabrication systems. Augmented reality was deployed to substantiate two-dimensional information by combining with digital fabrication equipment to enforce interaction. Fuzzy Delphi method and a two-dimensional quality questionnaire were used to consult expert opinions to optimize the model for increasing overall system satisfaction. The result of this project would be helpful for the exchange and diffusion of digital knowledge to strengthen user interaction with information. It could improve the efficiency of entire value chains including product development, marketing and service, which is very important to future innovative design and fabrication practice.

**AUTHORS/INSTITUTIONS:** C. Ko, Department of Design, National Taiwan University of Science and Technology, Taipei, TAIWAN|

**CONTROL ID:** 3378510

**TITLE:** Merging versus Ensembling in Multi-Study Machine Learning: Theoretical Insight from Random Effects

**ABSTRACT BODY:**

**Abstract Body:** A critical decision point when training predictors using multiple studies is whether these studies should be combined or treated separately. We compare two multi-study learning approaches in the presence of potential heterogeneity in predictor-outcome relationships across datasets. We consider 1) merging all of the datasets and training a single learner, and 2) cross-study learning, which involves training a separate learner on each dataset and combining the resulting predictions. In a linear regression setting, we show analytically and confirm via simulation that merging yields lower prediction error than cross-study learning when the predictor-outcome relationships are relatively homogeneous across studies. However, as heterogeneity increases, there exists a transition point beyond which cross-study learning outperforms merging. We provide an analytic expression for the transition point that can be used to help guide decisions about whether to merge data from multiple studies.

**AUTHORS/INSTITUTIONS:** Z. Guan, G. Parmigiani, Harvard University, Cambridge, Massachusetts, UNITED STATES|P. Patil, Boston University, Boston, Massachusetts, UNITED STATES|

**CONTROL ID:** 3378568

**TITLE:** A Method to Flexibly Incorporate Covariates in Latent Class Regression with Application to Mild Cognitive Impairment

**ABSTRACT BODY:**

**Abstract Body:** Latent class analysis is a powerful statistical method to elucidate the structure of a heterogeneous population. An important extension is latent class regression (Bandeem-Roche et al., 1997), which allows researchers to explore whether covariates are risk factors affecting the relative frequencies of the latent classes. For example, mild cognitive impairment (MCI) is a clinical construct representing a mixture of patient subpopulations with distinct underlying neuropathologies, including Alzheimer's disease and related neurodegenerative diseases, and there is considerable interest in investigating whether vascular covariates are risk factors for these unobserved disease subtypes. While latent class regression is potentially attractive in this application, we have found that entering covariates into the model can unintentionally alter the clinical interpretation of the latent classes of MCI. We introduce the concept of a covariate activity governor, which provides a flexible method for researchers to incorporate covariates without distorting too extensively the clinical conceptualization of the latent classes in the maximum likelihood solution. We apply our method to investigate the structure of MCI taking into account vascular covariates.

**AUTHORS/INSTITUTIONS:** J. Hanfelt, G. Kim, Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, UNITED STATES|F. Goldstein, Neurology, Emory University, Atlanta, Georgia, UNITED STATES|

**CONTROL ID:** 3379075

**TITLE:** Nonparametric inference procedures for panel count data with competing risks

**ABSTRACT BODY:**

**Abstract Body:** In survival and reliability studies, panel count data arise when we investigate a recurrent event process which is observed only at discrete time points. If each individual is exposed to the risk of recurrence of the event due to two or more different causes, we obtain competing risks panel count data. Such data arise frequently from transversal studies on recurring events in demography, epidemiology and reliability experiments where the individuals cannot be observed continuously. In the present paper, we propose an isotonic regression estimator for the cause specific mean function of the underlying recurrent event process of a competing risks panel count data. Further, a non-parametric test is proposed to compare the cause specific mean functions of the panel count competing risks data to test whether the effect of different risks on life time are identical or not. Asymptotic properties of the proposed estimator are studied which is used in deriving the test statistic. A simulation study is conducted to assess the finite sample behaviour of the proposed estimator and test statistic. The proposed method is demonstrated using a real life data arising from skin cancer chemo prevention trial given in Sun and Zhao (2013).

**AUTHORS/INSTITUTIONS:** S. Purushothaman, SNGSC, Pattambi, India, Pattambi, Palakkad, Kerala, INDIA|

**CONTROL ID:** 3379195

**TITLE:** Bias reduction and solution to separation in the AFT models for small or rare event survival data

**ABSTRACT BODY:**

**Abstract Body:** The Accelerated failure time (AFT) model is widely used in medical science and reliability engineering for its intuitive interpretation. The model parameters are generally estimated by maximum likelihood estimation (MLE) which reports unbiased and consistent estimates when sample size is large and/or rate of censoring is low; however, its small sample performance is unknown. This paper investigated the properties of MLEs of the regression parameters of the AFT models for small sample or rare-event (high rate of censoring) situation and introduced a penalized likelihood approach to address the problems. The penalized likelihood function and the corresponding score equation were derived by adding a penalty term, equivalent to the Jeffreys invariant prior, to the existing likelihood function, which was originally proposed by Firth (Biometrika, 1993) for reducing the first order bias in MLEs of the regression parameters of the exponential family models. The penalized method was illustrated for the most commonly applied log-location-scale family models such as Weibull, Log-normal and Log-logistic distributions. The illustration showed that the Jeffreys-prior based penalized likelihood succeeds to achieve convergence, providing finite estimates of the regression coefficients and solves the problem of separation or monotone likelihood created by a covariate, which are not often possible by the MLE. Extensive simulation studies conducted separately for each of the log-location-scale models demonstrated the penalized approach to have a substantial improvement over MLE by providing smaller amount of bias, mean squared error (MSE) and narrower confidence interval. An application of the methods using data with small sample and rare event supports the findings from the simulation. The penalized likelihood approach is therefore recommended to use in both cases of small ( $N < 50$ ) and large sample with high censoring rate in practice.

Keywords: bias reduction, monotone likelihood, Jeffreys prior, log-location-scale family.

**AUTHORS/INSTITUTIONS:** T. Alam, M.S. Rahman, Institute of Statistical Research and Training, University of Dhaka, Dhaka, BANGLADESH|

**CONTROL ID:** 3379236

**TITLE:** Semiparametric Transformation Model for Competing Risks Data with Cure Fraction

**ABSTRACT BODY:**

**Abstract Body:** We propose a new methods for analysing competing risks data with long term survivors. We formulate the effects of covariates on sub-survival functions of competing risks using linear transformation which enable us to estimate the cure fraction from the proposed model itself. We find the estimators of regression coefficients using counting process based estimating equations. The asymptotic properties of the estimators are studied using martingale theory. An extensive Monte Carlo simulation study is carried to asses the finite sample performance of the estimators. Finally, we illustrate our method using two real data sets.

**AUTHORS/INSTITUTIONS:** S.K. Kattumannil, Indian Statistical Institute, Chennai, Chennai, Tamil Nadu, INDIA|

**CONTROL ID:** 3380073

**TITLE:** Modeling HIV viral load rebound trajectories after antiretroviral treatment interruption

**ABSTRACT BODY:**

**Abstract Body:** Characterization of HIV viral rebound after the discontinuation of antiretroviral therapy is central to HIV cure research. We propose a parametric nonlinear mixed effects model for the viral rebound trajectory, which often has a rapid rise to a peak value and followed by a decrease to a viral load set point. We choose a flexible functional form that captures the shapes of viral rebound trajectories and can also provide biological insights regarding the rebound process. Each parameter can incorporate a random effect to allow for variation in parameters across individuals. Key features of viral rebound trajectories such as viral set points are represented by the parameters in the model, which facilitates assessment of covariate effects and identification of important pre-treatment interruption predictors for these features. We employ a stochastic expectation maximization algorithm to incorporate HIV-1 RNA values that are below the limit of assay quantification. We evaluate the performance of our model in simulation studies and applied the proposed model to longitudinal HIV-1 viral load data from five AIDS Clinical Trials Group treatment interruption studies.

**AUTHORS/INSTITUTIONS:** R. Wang, Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts, UNITED STATES|R. Wang, A. Bing, C. Wang, Y. Hu, R. Bosch, V. DeGruttola, Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, UNITED STATES|

**CONTROL ID:** 3380341

**TITLE:** Iterative independent component analysis for identifying cancer subtypes

**ABSTRACT BODY:**

**Abstract Body:** One of the important goals in cancer studies using omics data is to identify cancer subtypes. Since cancer is a complex disease caused by various etiological reasons, a strong heterogeneity always exist from patients to patients. Thus, in the precision medicine era subtyping homogeneous patients is very important to identify an appropriate treatment to each subtype of patients. While there have been several approaches proposed for subtyping, they mainly relied on simple dimensional reduction without considering hidden structures of data. We propose a novel a framework to identify cancer subtypes more efficiently through an iterative application of independent component analysis (ICA). ICA is an unsupervised approach to identifying and separating mixed sources from observed signals with little prior information. While principal component analysis (PCA) finds orthogonal components that follow a multivariate normal distribution, ICA identifies statistically independent components without a normality assumption. Unlike other existing methods, the proposed iterative ICA method take into account the hidden structures by independent components. Our approach consists of two stages: i) identifying cancer subtypes by iterative pruning via ICA and ii) predicting survival-risk labels. At the first stage, the dimension of original data is reduced by projecting them to a lower dimensional subspace built by ICA. The samples are then bisected. By repeating this procedure until a predetermined stopping criterion is satisfied, the subtype structures can be identified and survival-risk label is assigned to each subtype based on survival times. At the second stage, a classification model based on the support vector machine (SVM) is built to classify the subtypes derived from the first stage. We illustrate the proposed iterative ICA using the ovarian cancer data from The Cancer Genome Atlas (TCGA). The iterative ICA successfully identified ovarian cancer subtypes with quite different survival times.

**AUTHORS/INSTITUTIONS:** M. Park, Statistics, Eulji University, Daejeon, KOREA (THE REPUBLIC OF)|H. Lee, K. Moon, Statistics, Korea University, Seoul, KOREA (THE REPUBLIC OF)|T. Park, Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3380441

**TITLE:** A Comparison of Classification Methods for DNA-based Identification of Forensically Important Fly Species

**ABSTRACT BODY:**

**Abstract Body:** Forensic Entomology has been a powerful tool in the crime scene investigations over the past few years. Insects and other arthropods provide significant evidence to determine the time, manner or location of death. There are a variety of species which can be found on the corpse. Each species has different biological and morphological distinctions, and these distinctions can highly affect calculating minimum post-mortem interval or other crucial factors such as growth rate and development rate. Therefore, correct species identification is the primary and the most important step of forensic entomology. Traditionally, morphology-based approach has been widely used in the past. However, it is time-consuming and there is a limitation when no suitable specimen for morphological identification is obtained. Hence, the molecular identification is considered as an alternative method and it is broadly used nowadays. The cytochrome oxidase I (COI) as the marker for animal species discrimination has been shown to be useful for identification in many studies. Phylogenetic analysis is commonly used in forensic entomology, but classification using DNA sequencing was relatively few. After COI sequences are correctly amplified and sequenced, statistical classification analysis is conducted to identify species of blowflies.

The aim of this study is to contribute to the development and accuracy improvement of species classification of forensically important flies using statistical classification method. In this study, over 20 species were used to analyse the identification of fly species. Various statistical classification methods such as CART(Classification and Regression tree), C5.0, NB(Naïve Bayes), KNN(K nearest Neighbours), Bagging and xgboost(Extreme Gradient Boosting) were used to classify each species with DNA sequencing. Model evaluation criterion was the misclassification error rate, and we tried to find the best classification method with the lowest misclassification error rate.

**AUTHORS/INSTITUTIONS:** J. PARK, S. Jeong, J. Lee, Department of Statistics, Korea University, Seoul, KOREA (THE REPUBLIC OF)|S. Park, S. Shin, Korea Institute for Legal Medicine, Korea Univesity, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3380815

**TITLE:** Modelling the spread of an infectious disease using a spatio-stochastic SIR model.

**ABSTRACT BODY:**

**Abstract Body:** The spread of infectious diseases is a world-wide problem that has a greater and more damning impact on low-income countries. Mathematical modelling is a useful tool to better understand these diseases and to plan prevention and interventions. This article extends the simple stochastic epidemic model derived by Tuckwell and Williams (2007) by the addition of a spatial component. The stochastic SIR model is first defined and then the spatial component is included in the function which represents the number of contacts that the individual makes. Simulations are then used to compare the models with and without the spatial component. The addition of a spatial component makes intuitive sense as infectious diseases are spread in relation to the spatial structure of the population.

**References:**

Tuckwell, H. C., & Williams, R. J. (2007). Some properties of a simple stochastic epidemic model of SIR type. *Mathematical biosciences*, 208(1), 76-97.

**AUTHORS/INSTITUTIONS:** R. Manjoo, School of Statistics and Actuarial Sciences, University of Witwatersrand, Johannesburg, SOUTH AFRICA]

**CONTROL ID:** 3381255

**TITLE:** Evaluating concordance between automated and manual scoring of Polysomnographic Recordings from a clinical trial using zolpidem in the treatment of insomnia

**ABSTRACT BODY:**

**Abstract Body:** When we get to compare instruments (methods or processes) in the field of health, we need to select the instrument, which is less valuable, easy to use, and less invasive one. Usually, in trial clinical to take the best decision in order to exchange the instruments, we have to produce reliable and valid measurements. Nevertheless, in contemplation of getting to the right choice, how to do it?

Svetnik et al. (2007) conducted a clinical study designed to compare the automated and semi-automated scoring of Polysomnographic (PSG) recordings used to diagnose transient sleep disorders. The study included 82 patients who were given a sleep-inducing drug (Zolpidem 10 mg). Measurements of latency to persistent sleep (LPS: lights out to the beginning of 10 consecutive minutes of uninterrupted sleep) were obtained using six different methods. Out of the six considered processes in the original study, in this scenario only two were selected. The first is manual scoring (Manual) and the second, is automated scoring by the Morpheus software (Automatic) to evaluate the concordance between measurements of two alternative methods with the concordance correlation coefficient (CCC) proposed for Lin (1989).

When moving from design-based estimation methods to model-based estimation methods, the evaluation of model assumptions are important. Evaluating the degree of agreement between estimates derived from these alternative methods also seems like a desirable property to assess. Especially when new estimation methods are to be introduced in replacement of methods that are already in place for deriving the target estimates. In practice it is common to assume that the data is normally distributed, however in the presence of outliers, the normality assumption is no larger valid, leading to biased estimates of the CCC, thus affecting the decision about the agreement between the measurements. To overcome the above problem, we propose to estimate the CCC based on the t-Student distribution. Further, to detect the sensitivity at the estimator of the CCC, different perturbations scheme are applied and local influence is studied.

**AUTHORS/INSTITUTIONS:** C.G. Leal Kaymaliz, Escuela de Salud, Universidad Viña del Mar, Viña del Mar, Valparaíso, CHILE|M. Galea Rojas, Departamento de Estadística, Pontificia Universidad Católica de Chile, Macul, Santiago, CHILE|F. Osorio, Departamento de Matemática, Universidad Técnica Federico Santa María, Valparaíso, Valparaíso, CHILE|

**CONTROL ID:** 3382610

**TITLE:** A comparison of two estimators of the treatment difference for phase II/III clinical trials with a control

**ABSTRACT BODY:**

**Abstract Body:** A seamless phase II/III clinical trial compares a number of drugs or doses with the control treatment in a single trial conducted in two stages. The first stage studies all of the experimental treatments and selects the one with the largest sample mean. This selected treatment along with the control treatment will continue to the second stage for further analysis. A treatment has to perform well in the first stage to be selected and continue to the second stage. Therefore, the sample mean for the selected treatment is a positively biased estimator of the corresponding population mean. Since the maximum likelihood estimator is also biased, due to combining data from both stages, a uniformly minimum variance conditionally unbiased estimator for the difference between the selected and control treatments has been proposed when two experimental treatments and a control are compared in the first stage. A comparison of the properties and distributions of these two estimators has been made by obtaining the exact expectation, bias and variance of the maximum likelihood estimator, and then considering special cases where this estimator is as good as the conditionally unbiased estimator. The assumption is that the treatment responses have normal distributions with unequal known variances and there is unequal allocation of patients. To illustrate the results, simulation studies have been carried out using R. The results show that the bias of the maximum likelihood estimator increases when there are more treatments initially. Furthermore, this estimator is unbiased when the variances are equal and there is equal allocation of patients in the first stage. Although the bias of the maximum likelihood estimator can be substantial, it has a lower mean square error than the conditionally unbiased estimator for all futility boundary values. Moreover, the mean square error of the latter estimator decreases as the difference in the means, both the best and second best with the mean of the control, increases and seems to approach the mean square error of the maximum likelihood estimator. Overall, the distributions of the two estimators are very similar when the means of the best and the second best treatments are significantly larger than or smaller than the mean of the control treatment for a given value of the futility boundary.

**AUTHORS/INSTITUTIONS:** A.J. Mohammad, School of Mathematical Sciences, Queen Mary University of London, London, UK, UNITED KINGDOM|

**CONTROL ID:** 3383433

**TITLE:** Prediction performance with correlated data

**ABSTRACT BODY:**

**Abstract Body:** Generalized linear mixed models (GLMM) have grown in their use in the practice of biometric analysis, as they enable answering important research questions with data from complex hierarchical study designs, when observations are not independent. When using these models, one may be interested in exploring associations between the outcome variable with some potential factors, or in predicting the outcome given certain factors. In the so-called 'association models,' models are judged based on the interpretation of the regression coefficients, while in the so-called 'prediction models,' models are judged by comparing the observed values with the model predicted values. The statistical literature does not describe well how to handle the random effects when predicting the outcome after fitting GLMMs. This article aims to compare the predictions constructed from a logistic regression mixed effects model with and without random effects, and those from naïve logistic regression models. Specifically, it investigates whether to consider or not the random effects in the predictions. We focus on the GLMM when the dependent variable is binary, since it is a commonly used model in biometric practice. The indicators used to judge the predictions include the area under the receiver operating characteristic (ROC) curve (AUC), the Bangdiwala B statistic for concordance, and the proportion of misclassification.

The insights and resulting guidance from this work stem from two approaches: (1) simulated data scenarios that cover the gamut of real-life situations, and (2) real data cross-validation analyses. In (1), correlated data were simulated using the R package SimCorMultRes (Touloumis, 2017) for different values of the model parameters, degrees of correlation between observations, size and number of clusters considered. In all situations, we found that GLMM predictions that include random effects are overwhelmingly better than the GLMM predictions that do not consider the random effects, which resemble the naïve logistic regression predictions, and thus the incorrect model for clustered correlated data. In (2), we illustrate the issues using real data that motivated exploring this aspect of statistical practice. The results of this application are consistent with those obtained in the simulation study.

We thus recommend that when obtaining predictions, researchers should use the random effects in their predictions.

**AUTHORS/INSTITUTIONS:** S. Bangdiwala, Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, CANADA|S. Bangdiwala, Institute for Health and Social Sciences, University of South Africa, Lenasia, SOUTH AFRICA|A.M. Sfer, Universidad Nacional de Tucuman, Tucuman, ARGENTINA|M.A. D'Urso Villar, Investigación, Universidad Nacional de Tucumán, Yerba Buena, Tucumán, ARGENTINA|

**CONTROL ID:** 3383689

**TITLE:** Evaluation of prediction performance for Cox proportional hazards model

Masaaki Matsuura, Tasuku Yoshida (Teikyo University Graduate School) and Ando Shuji (Tokyo university of Science)

**ABSTRACT BODY:**

**Abstract Body:** Abstract: The Cox proportional hazards model is widely used and very useful for evaluation of some variables. We introduce an evaluation method of prediction performance by using set of estimated regression coefficients by the Cox models. This evaluation is an extension of the method using ROC curves based on the results from logistic regression model. For illustrative example, we will give a result from real data analysis of risk factors related to readmission to an intermediate care unit (IMCU) in a hospital.

**Introduction and method:** We can estimate probabilities of event of interest for each subject in a cohort analysis by the estimated regression coefficients from logistic regression model. Furthermore we can calculate accuracy, sensitivity and specificity based on the ROC curve for the estimated probabilities and the true status of event's result. Using similar manner, we can examine prediction performances based on the Cox proportional hazards model. The method is based on the use of logistic transformation of a value  $Z$  defined by a linear combination of estimated regression coefficients and covariate-values by the used Cox model. We used the value  $Z$  for the Cox model instead of the estimated probabilities by the logistic regression model.

**Example:** We used a retrospective cohort data for an analysis of risk factors related to re-enter a IMCU in a hospital. The study period is from April 1, 2016 to May 31, 2019. The total subjects are 1633 patients, of which 88 patients (5.4%) had experiences of readmission to an intermediate care unit (IMCU). We will compare both prediction performance results from a Cox model and a logistic model at the meeting.

**Discussion:** Survival data are incomplete data and censoring mechanisms differs for each other. The performance of prediction based on the Cox model may depend on the censoring mechanisms. We will discuss this problems in the meeting.

**AUTHORS/INSTITUTIONS:** M. Matsuura, T. Yoshida, Teikyo University Graduate School of Public Health, Itabashi-ku, Tokyo, JAPAN|S. Ando, Tokyo University of Science, Tokyo, JAPAN|

**CONTROL ID:** 3383764

**TITLE:** Under-five mortality in India: An application of multilevel cox proportional hazard model

**ABSTRACT BODY:**

**Abstract Body:** Although consideration work has been done to understand the effect of individual level factors on under-five mortality, less is known about the community (neighbourhood) characteristics affect health outcomes for children, even though they have a prominent role in theoretical model. This study address important issues in under-five mortality in India. The objective of this paper is to determine the important of community, household and individual level effect on under-five mortality in India. Using data from the latest round of Demographic Health Survey (DHS)-2015-16, known as National Family Health Survey (NFHS) in India, multilevel cox proportional hazard analysis was performed on a nationally representative sample of 259,627 children nested within 180,227 household who were also nested within 28332 communities. Hazard ratio (HR) with 95% confidence interval (CI) were used to express measure of association among the characteristics. Variance partition coefficient (VPC) and Wald statistics were used to express measures of variation. The results indicate that pattern of under-five mortality were clustered within household and communities. The community level variables like region, place of residence, community poverty level, community education level, ethnic fractionalization index were significantly determine under-five mortality in India. The risk of under-five deaths were significantly higher for children residing in North (HR: 1.34; 95% CI: 1.15-1.60), East (HR: 1.84; 95% CI: 1.59-2.13) and West regions (HR: 1.45; 95% CI: 1.23-1.70) compared to South region. In addition, the proportion of women in community completing secondary school (HR: 0.58; 95% CI: 0.51-0.66) were significantly more likely to increase the child survival. The mother level variables like maternal education, BMI, mother age at birth and breast-feeding were significantly determine under-five mortality. The results suggest to address the contextual level factors to address under-five mortality in India. The findings also suggest the need to focus on community-level intervention aim at improving the socioeconomic conditions of mothers, especially disadvantage regions such as North, East and West.

**AUTHORS/INSTITUTIONS:** A. Yadav, Department of Mathematical Demography and Statistics , International Institute for Population sciences, Noida, Uttar Pradesh, INDIA|

**CONTROL ID:** 3383809

**TITLE:** Joint Penalized Smoothing Spline Modeling of Multivariate Longitudinal Data, with Application to HIV-1 RNA Load Levels and CD4 Cell Counts

**ABSTRACT BODY:**

**Abstract Body:** Motivated by the need to jointly model the longitudinal trajectories of HIV viral load levels and CD4 counts during the primary infection stage, we propose a joint penalized smoothing spline modeling approach that can be used to model the repeated measurements from multiple biomarkers of various types (e.g., continuous, binary) simultaneously. This approach allows for flexible trajectories for each marker, accounts for potentially time-varying correlation between markers, and is robust to mis-specification of knots. Despite its advantages, the application of multivariate penalized smoothing spline models, especially when biomarkers may be of different data types, has been limited in part due to its seemingly complex implementation. To overcome this, we describe a procedure that transforms the multivariate setting to the univariate one, and then makes use of the generalized linear mixed effect model representation of a penalized smoothing spline model to facilitate its implementation with standard statistical software. We performed simulation studies to evaluate the validity and efficiency through joint modeling of correlated biomarkers measured longitudinally compared to the univariate modeling approach. We applied this modeling approach to longitudinal HIV-1 RNA load and CD4 count data from Southern African cohorts to estimate features of the joint distributions such as the correlation and the proportion of subjects with high viral load and high CD4 count over time.

**AUTHORS/INSTITUTIONS:** T. Chen, R. Wang, Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts, UNITED STATES|L. Zhao, Department of Prevention Medicine, Northwestern University, Chicago, Illinois, UNITED STATES|V. Novitsky, Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, UNITED STATES|R. Wang, Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, UNITED STATES|

**CONTROL ID:** 3383839

**TITLE:** Permutation-based Variable Importance Measures for Unsupervised Random Forests

**ABSTRACT BODY:**

**Abstract Body:** Random Forests (RF) have been shown to be fast and to produce good results in many high-dimensional applications such as genome-wide association and gene expression microarray studies. In addition to using supervised RF to solve classification problems, RF permutation-based Variable Importance Measures (VIMs) are frequently used for features selection to differentiate variable signal from noise, and many permutation approaches have been proposed for this purpose. These are expected not only to clearly differentiate signal from noise, but also, for noise variables, to have a distribution of importance estimates centered around zero. In principle, permutation-based VIMs can also be estimated from unsupervised RF, where no target variable is observed but artificially generated based on the marginal distributions of predictors. In this case, the distribution of the VIM could be affected by the marginal distribution of predictors. Our work investigates different proposed VIMs in the case of unsupervised RF, presents their limitations, and proposes a new permutation based VIM inspired by cross-validation and the holdout trick as suggested in the literature. Our new approach is expected to be fast and produce better results in terms of recognizing signal as well as to have importance estimates of noise variables centered around zero for both cases of supervised and unsupervised RF. The proposed approach will be evaluated on different artificial and real datasets.

**AUTHORS/INSTITUTIONS:** C.J. Fouodo, I.R. König, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany, Luebeck, GERMANY|

**CONTROL ID:** 3383855

**TITLE:** Designing pharmacokinetic experiments in the presence of competing covariates

**ABSTRACT BODY:**

**Abstract Body:** In pre-clinical stage of drug development process, pharmacokinetic experiments are conducted to examine the association between substrate concentrations and the enzymes that could be responsible for metabolism of the substrate. Each of such experiments is conducted using a level of substrate concentration and a sample of a human liver microsome (HLM), and the rate of the reaction is measured. Such experiment is repeated for different combinations of substrate concentrations and HLM samples. HLMs are different from each other in terms of its affinity to different enzymes and the strength of affinity (activity values) of different enzymes are available for each HLM at the design stage. Since a number of experiments are performed with the same HLM and the rate of reactions measured from the experiments with a specific HLM are assumed to be correlated. A transform-both-sides nonlinear (e.g. Michealis-Menten model) mixed effects models can be considered to quantify the association between the substrate concentration and related enzymes, where model parameters are assumed to be different for different HLMs used in the experiments. Depending on the substrate, one or more enzymes could be responsible for its metabolism. Optimum design of such pharmacokinetic experiments deals with selecting  $n$ , the pre-specified number of experiments to be conducted, pairs of concentration and HLM sample to be used in the experiments, so that the resulting conclusion could be the most efficient. In this study, a number of design criteria are proposed to obtain exact optimum pharmacokinetic experiments. Exact optimum designs of different number of runs are obtained for both single-enzyme and multiple-enzyme models, and the performance of calculated optimum designs are compared with the rich design, a design that is based on all possible pairs of substrate concentrations and available HLMs. A number of simulation studies are performed and its results show that the optimum designs with a fewer number of runs can outperform the rich design in providing efficient estimates of model parameters.

**AUTHORS/INSTITUTIONS:** B. Bogacka, Department of Mathematical Sciences, Queen Mary University of London, London, UNITED KINGDOM|M.A. Latif, Institute of Statistical Research and Training, University of Dhaka, Dhaka, BANGLADESH|S. Gilmour, Department of Mathematics, King's College London, London, UNITED KINGDOM|

**CONTROL ID:** 3383885

**TITLE:** Modelling Herbal Drug Mixture for Diabetes Mellitus Using Response Surface Methodology

**ABSTRACT BODY:**

**Abstract Body:** The research on which this paper is based modelled a herbal drug mixture of diabetes mellitus using response surface methodology. In the study, three herbs commonly used in Kenyan communities as anti-hyperglycaemic drugs were investigated. The herbs tested were *Moringa oleifera*, *Urtica dioica* (stinging nettle) and *Cinnamomum verum* (cinnamon). A simplex-centroid design for the three herbal drug mixture components was set up as an experiment in the laboratory. A total of 7 herbal mixtures were tested using 7 groups of 4 alloxan- induced diabetic female albino wistar rats with a body weight between 110-140g. Two control groups, each with four diabetic rats, were used as follows: group 8 considered a positive control, was given conventional diabetic medicine (metformin) and group 9 was considered as a negative control and was given distilled water. The response measured was the drop in Blood Sugar Level (BSL) between the time of administration of the drug, considered as zero hour and two hours after using two different dose rate levels 250 mg/kg and 500mg/kg of the herbal mixture. All the treatments were in aqueous form and were orally administered through Gavages' method. Scheffé mixture models were fitted to the experimental data and the analysis done in the statistical environment of R software. At 250mg/kg dosage the results showed that using a quadratic model to describe the surface of the changes in BSL is not justified; instead, the linear model is adequate. On the other hand, at 500mg/kg dosage level a quadratic model was justified. A mixture of *Urtica dioica* and *Cinnamomum verum* was highly synergistic and it was even more effective than any of the pure herbal components at the 500mg/kg. The mixture of *Moringa oleifera*, *Urtica dioica* and *Cinnamomum verum* were antagonistic. The model p-values were significant for both linear and quadratic models. The parameter estimates for the single herbs were all statistically significant while the binary and tertiary mixture components were not statistically significant at  $\alpha=0.05$  level of significance. From the research study, it is recommended that pure herbs be used especially *Urtica dioica* at low dosage levels while a mixture of *Urtica dioica* and *Cinnamomum verum* can be used at higher dose.

**AUTHORS/INSTITUTIONS:** G.G. Njoroge, Physical Science , Chuka University, Nairobi, KENYA|

**CONTROL ID:** 3384143

**TITLE:** A statistical tolerance region for a heterogeneous urban wind field based on a small network of weather stations

**ABSTRACT BODY:**

**Abstract Body:** The vectorial wind field is a major factor in the transport and dispersion of air pollution in urban regions. This field is frequently characterized by an inherent spatial heterogeneity. This heterogeneity may be manifested by noticeable differences between rooftop level measurements in adjacent locations. Quite often the degree of heterogeneity changes through the day.

A possible way to obtain real-time information on the current state of the urban wind field is to use a dense weather stations' network. Such endeavor is expansive and technically demanding. This leads to networks that are too sparse to accurately describe the changing degree of the urban wind vectors' heterogeneity.

In situations where there is not sufficient information regarding a specific urban region, it is possible to conduct a two phase scheme. First, deploy a dense weather stations network for a limited time. Then leave a sparse network to continuously monitor the region. Such scheme requires a statistical estimation method that will use the sparse network's measurements to compensate for the missing information.

A useful estimation for the degree of the wind field's heterogeneity is a tolerance region that contains a given proportion of the wind vectors' population. For vectors that follow the bivariate normal distribution, an ellipse shaped tolerance region can be constructed. This tolerance region uses an analytical Mahalanobis distance function, which is based on the F distribution. However, because actual measurements of urban wind vectors does not always follow this distribution, another approach is required.

This study used wind data collected in the metropolitan area of Tel-Aviv. The results show that the spatial wind distribution can be very well represented by a small sample of merely four stations. Based on this sample stations, empirical Mahalanobis distance functions were calculated for each season. These functions were found to fit well the logistic distribution. These functions were validated by applying them for tolerance regions on a different data set.

**AUTHORS/INSTITUTIONS:** Z. Klausner, E. Fattal, Applied Math, Israel Institute for Biological Research, Ness Ziona, ISRAEL|

**CONTROL ID:** 3384248

**TITLE:** A decision-theoretic approach to Bayesian clinical trial design and evaluation of robustness to prior-data conflict

**ABSTRACT BODY:**

**Abstract Body:** Bayesian clinical trials allow taking advantage of relevant external information through the elicitation of prior distributions, which influence Bayesian posterior parameter estimates and test decisions. The impact of prior specification on frequentist (conditional) operating characteristics is generally investigated at the design stage of the trial. However, as power gains are typically not possible when requiring strict type I error probability control [1], it is of interest to investigate principled approaches to relax such requirement when borrowing is desired.

Here we approach this task from a Bayesian decision theoretic standpoint. Recent proposals targeting Bayesian (average) errors (see e.g. [2, 3]) are reviewed and incorporated in a global approach which additionally takes into account the relative costs of estimation error and of sampling. Sensitivity analyses are performed by making a distinction between the prior of the data-generating process, and the analysis prior adopted to fit the data, a distinction adopted also in e.g. [4]. We demonstrate the applicability of the approach through simulations, where robustification approaches such as mixture and empirical Bayes power analysis prior specifications are included for comparison.

[1] Kopp-Schneider, A., Calderazzo, S., & Wiesenfarth, M., *Biometrical Journal*, 2019.

[2] Pericchi, L., & Pereira, C, *Brazilian Journal of Probability and Statistics*, 2016.

[3] Psioda, M. A., & Ibrahim, J. G., *Biostatistics*, 2018.

[4] Sahu, S.K., & Smith, T.M.F., *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2006.

**AUTHORS/INSTITUTIONS:** S. Calderazzo, Biostatistics, German Cancer Research Center, Heidelberg, GERMANY|M. Wiesenfarth, German Cancer Research Center, Heidelberg, GERMANY|A. Kopp-Schneider, Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, GERMANY|

**CONTROL ID:** 3384394

**TITLE:** Surrogate marker evaluation from a clinical trial with continuous longitudinal surrogate endpoints

**ABSTRACT BODY:**

**Abstract Body:** Evaluation of a new medical intervention(Z) should be based on the most clinically relevant endpoints (T), which could be rare and thus require a large sample size or follow patients for a long time. Some other clinical endpoints and surrogate biomarkers (S) are observed earlier, easier, possible repeated, and cost-saving can be used to surrogate T in clinical trial research. Surrogate biomarkers (S) follows a stochastic process during the trial. The scientific challenge is to transfer the observed treatment effect from S to T . In this paper, we use the information-theoretic measure of association (ITMA) for surrogacy based on Havrda-Charva (HC) entropy. A general model with three kinds of special cases is given in formulas. We derive the estimation procedures to calculate these entropy functions and demonstrate how to use the HC entropy to derive a more normally distributed estimate of ITMA for surrogacy. We also show the use of ITMA for surrogacy to guide design for the number of visits in clinical trials. Real application examples and simulation studies are discussed with the new models.

**AUTHORS/INSTITUTIONS:** Q. ZHAO, School of Public Health, Guangzhou Medical University, Guangzhou, CHINA|Q. ZHAO, Y. Lu, Stanford University, Palo Alto, California, UNITED STATES|M.D. Pardo, Complutense University of Madrid, Madrid, SPAIN|J. Hua, South China Normal University , Guangzhou, CHINA|

**CONTROL ID:** 3384511

**TITLE:** A Cluster Based Sparse Canonical Correlation Analysis to Test Associations Between a Multi-Pollutant Mixture and High-Dimensional DNA Methylation

**ABSTRACT BODY:**

**Abstract Body:** Emerging studies show that exposure to fine particulate matter (PM<sub>2.5</sub>) may alter epigenetic traits such as DNA methylation. Understanding the toxic elements among components of PM<sub>2.5</sub> is highly important in order to provide the effectiveness of pollution controls. However, there is lack of statistical methods that rigorously quantify pollution effects in high-dimensional data such as epigenome data. Moreover, since the concentrations of these components of PM<sub>2.5</sub> are highly correlated due to its same emission source, it is extremely difficult to disentangle their independent effects.

Existing methods such as site-by-site testing procedures may lead to underpowered analysis, which can occur especially in the presence of weak signals of exposure effects. Recent studies show that power of the analysis can be increased by combining information across adjacent sites, as neighboring methylation sites (e.g. CpG) tends to be correlated. Although this may lead to power gain, the analysis is highly challenging due to 1) the highly correlated multiple exposures, 2) correlation among CpG sites, and 3) high dimensional data.

We propose a cluster based Sparse Canonical Correlation Analysis (sCCA) to test associations between a multi-pollutant mixture and high-dimensional DNA methylation. We first apply a clustering algorithm that clusters correlated neighboring CpG sites, and then analyze the association between each cluster of regions and PM<sub>2.5</sub> species concentrations using sCCA, which applies a penalty function to perform model selection. We compare a variety of penalty functions in the sCCA. We evaluate the performance of this two-stage analysis when using a variety of penalty functions in the sCCA. Through simulation studies, we show that the proposed method yields greater efficiency relative to existing methods.

**AUTHORS/INSTITUTIONS:** J.J. Lee, T. Sofer, B.A. Coull, Department of Biostatistics, Harvard Chan School of Public Health, Boston, Massachusetts, UNITED STATES|T. Sofer, Department of Medicine, Harvard Medical School, Boston, Massachusetts, UNITED STATES|T. Sofer, Department of Sleep Medicine, Brigham and Women's Hospital, Boston, Massachusetts, UNITED STATES|

**CONTROL ID:** 3384548

**TITLE:** Moment-based Estimation of Mixtures of Regression Models

**ABSTRACT BODY:**

**Abstract Body:** Finite mixtures of regression models provide a flexible modeling framework for many phenomena. Using moment-based estimation of the regression parameters, we develop unbiased estimators with a minimum of assumptions on the mixture components. Our proposed method imposes no assumptions on the distributions of the individual components and only a minimum of restrictions on the regression effects. In particular, only the average regression model for one of the components in the mixture model needs to be specified. We derive consistency and asymptotic distribution of the estimators, show how to use them for hypothesis testing and show how moment-based mixture of regression models estimators that can be applied to a large number of situations.

Since computation is fast the method is applicable for large-scale studies such as genome-wide association studies and we illustrate the use of the moment-based mixture of regression models with an application in genomic data analysis where we search for gene-environment interactions of single-nucleotide-polymorphisms (SNPs) that were undiscovered using traditional approaches.

**AUTHORS/INSTITUTIONS:** C. Ekstrom, Biostatistics, University of Copenhagen, Copenhagen, Copenhagen, DENMARK|C. Pippert, Leo Pharma, Ballerup, DENMARK|

**CONTROL ID:** 3384892

**TITLE:** Genetic correlation analysis of liver cancer using multi-traits from UK Biobank

**ABSTRACT BODY:**

**Abstract Body:** Estimating genetic correlations between complex phenotypes can provide valuable insights into etiologic studies. Major challenge for estimating genetic correlation from genome-wide association studies (GWAS) is the insufficient availability of individual-level genotype data and sample overlap among meta-analyses. Linkage disequilibrium (LD) score regression using GWAS summary statistics allow us to quantify the genetic correlation and SNP heritability between pairs of traits. This approach could also provide insights into disease co-development potentially associated with disease risk.

We calculated genetic correlation ( $r_g$ ) between hepatocellular carcinoma (HCC) from Hepatocellular Carcinoma Epidemiology Consortium and other phenotypes from UK Biobank, a large cohort study enrolled over 500,000 adults of age 40-69 years in 2006-2010. .

We identified several phenotypes in blood counts, metabolic traits, alcohol consumption, smoking, diabetes, family history, and BMI associated with HCC at the significance level of  $P < 5 \times 10^{-2}$ . Alcohol intake frequency shows genetic correlation with HCC ( $r_g = 0.257$ ,  $P = 4.03 \times 10^{-4}$ ). Both HDL cholesterol ( $r_g = -0.251$ ,  $P = 3.77 \times 10^{-4}$ ) and apolipoprotein ( $r_g = -0.221$ ,  $P = 1.48 \times 10^{-3}$ ) in the blood were negatively correlated with HCC while both C-reactive protein ( $r_g = 0.268$ ,  $P = 7.50 \times 10^{-4}$ ) and haemoglobin concentration ( $r_g = 0.186$ ,  $P = 8.58 \times 10^{-4}$ ) were positively correlated. Anaemia due to iron deficiency showed the strongest correlation ( $r_g = 0.478$ ,  $P = 3.11 \times 10^{-2}$ ). Diabetes showed the positive association ( $r_g = 0.407$ ,  $P = 7.53 \times 10^{-5}$ ). Family history of diabetes in mother ( $r_g = 0.365$ ,  $P = 4.94 \times 10^{-4}$ ) and in siblings ( $r_g = 0.420$ ,  $P = 8.24 \times 10^{-4}$ ) was positively correlated.

LD score regression analysis provides an improved understanding of the genetic architecture between HCC and biomarkers. Mendelian randomization analyses can validate the potential causal direction between risk of HCC and phenotypes of interest.

**AUTHORS/INSTITUTIONS:** Y. Han, J. Byun, C.I. Amos, Medicine:Epidemiology, Baylor College of Medicine, Houston, Texas, UNITED STATES|Y. Han, C. Zhu, J. Byun, C.I. Amos, Institute for clinical and translational research, Baylor College of Medicine, Houston, Texas, UNITED STATES|D. Li, M. Hassan, The University of Texas MD Anderson Cancer Center, Houston, Texas, UNITED STATES|

**CONTROL ID:** 3384929

**TITLE:** Structural penalization in time-dependent Cox models and an application in mortality prediction for patients taking anticoagulants after atrial fibrillation.

**ABSTRACT BODY:**

**Abstract Body:** When dealing with a large number of candidate variables, variable selection is often necessary to establish a parsimonious model for prediction. We address two potential complications which arise in time-dependent Cox models to predict time to event. First, the candidate variables may change over time. Second, there may be an inherent grouping structure (GS) or strong/weak heredity among these variables. For example, 1) the several binary indicators representing a single categorical variable should be collectively included or excluded, 2) interaction selection is dependent on the selection of the main terms. Our proposed technique uses sparsity-inducing penalties on groups of variables, permitting user-specified GS incorporating a priori knowledge about the relationships among candidate variables and their interactions. Such dependence can also be nested, which calls for multiple penalty terms. Optimization is performed using a hybrid of accelerated proximal gradient descent with blockwise coordinate descent to improve efficiency. We further apply our method to a large dataset. Recently, 33,000 patients taking oral anticoagulants (OACs) for non-valvular atrial fibrillation and other drugs for comorbidities were identified from electronic health records in Quebec, Canada and followed up for one year. The outcome of interest is time to death. More than 50 time-varying and 12 baseline covariates were collected. Our goal is to build a parsimonious explainable predictive model for mortality and simultaneously understand the relationships among the variables and outcome, and thus inform clinicians of the factors that are most predictive and patients most at-risk. In terms of GS, the prescription and dose level of a particular OAC precede the adherence level; it is only meaningful to consider adherence when the OAC is prescribed. Furthermore, the GS becomes challenging when there are interactions between dose and adherence level of OACs, and drug interaction between OACs and other drugs such as antiplatelet and antidepressant when the OACs' dose and adherence level were available. After making a rational GS, can we select which potential variables are relevant in predicting mortality. We therefore provide several possible GSs based on scientific interest, then select variables that best predict mortality.

**AUTHORS/INSTITUTIONS:** G. Wang, R. Platt, Department of Epidemiology, Biostatistics and Occupation Health, McGill University, Montreal, Quebec, CANADA|R. Wang, Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts, UNITED STATES|R. Wang, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, UNITED STATES|T. Chen, Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts, UNITED STATES|M. Schnitzer, S. Perreault, Faculty of Pharmacy and School of Public Health, University of Montreal, Montreal, Quebec, CANADA|Y. Yang, Department of Mathematics and Statistics, McGill University, Montreal, Quebec, CANADA|

**CONTROL ID:** 3385425

**TITLE:**

SENSITIVITY OF GENOTYPE BY ENVIRONMENT INTERACTION MODELS TO OUTLYING OBSERVATIONS

**ABSTRACT BODY:**

**Abstract Body:** Plant breeding program depends on its ability to provide farmers with genotypes with guaranteed superior performance (phenotype) in terms of yield and/or quality across a range of environmental conditions. To achieve this aim, it is necessary to have an understanding of the model suitable for or leading to a good phenotype. In this study, two cases of scenarios were considered to have a clearer view of the performance of Genotype by Environment Interaction on the following four models; AMMI, FW, GGE and Mixed Model. We experiment the inference behind the violation of the assumption of normal distribution by observing the data contamination of two case scenarios (Lowest and Highest outlying observations). It was observed on the two data Types of Balance and Unbalance designs with different Levels of generations. We achieved that by comparative performance of the data contamination techniques under the two case scenarios; Case I scenario was done for Lowest Outlying Observations where 50%, 100% and 500% data contamination on the First Quarter (P1), Mid quarter (P2) and Last Quarter (P3). We then deduced from the result of the model evaluation that, at each levels of data contamination for Balance and Unbalance design, Mixed model was the ideal model for G by E interaction. Case II scenario was done for Highest Outlying Observations where 50%, 100% and 500% data contamination on the First Quarter (P1), Mid quarter (P2) and Last Quarter (P3) were examined on each levels of generations. We then observed from the result of the model evaluation that, at each levels of data contamination for Balance and Unbalance design, Mixed model also outperformed the other three models.

**AUTHORS/INSTITUTIONS:** O.S. Oyamakin, Statistics, University of Ibadan, Nigeria, Ibadan, Oyo, NIGERIA|M.O. Durojaiye, Department of Statistics, University of Ibadan, Ibadan, Oyo State, NIGERIA|

**CONTROL ID:** 3385445

**TITLE:** Weighted Composite Time-To-Event Endpoints with Recurrent Events: Comparing Three Analytical Approaches

**ABSTRACT BODY:**

**Abstract Body:** In many clinical studies the interest lies in the comparison of a treatment to a control regarding a time-to-event endpoint like time to myocardial infarction or time to death. Commonly, only one of those endpoints is considered in an analysis. A composite endpoint is an alternative approach where endpoints can be considered jointly. Usually the time to the first occurring event for an individual is thereby analyzed. However, an individual may experience more than one non-fatal event. By including all observed events in an analysis the effect estimates are based on more complete information. Thus, analytical methods for recurrent events are of interest where several event types, often of different clinical relevance, are considered. In such a case, weighting the event types regarding their clinical relevance was proposed. Such weight-based methods include the Wei-Lachin multivariate procedure, the weighted hazards approach, and Bakal's weighted composite endpoint. There exists no systematic comparison of these methods.

We therefore provide a simulation-based comparison of methods for weighted composite endpoints combining events of different clinical relevance; one recurrent non-fatal and one fatal. The methods are also applied to data from the GENESIS ('Genetic, Socio-economic and Inflammatory Determinants of Ischemic Stroke and their Interdependence') study conducted in Ludwigshafen, Germany.

For all approaches a closed formula test statistic is provided based on similar assumptions but different modeling ideas. For the Wei-Lachin approach and the weighted hazards method a weighted effect measure is described but not for Bakal's approach. Confidence intervals can only be easily gained for the Wei-Lachin effect measure. Differences in the empirical power can be seen as well as a smaller mean squared error for the weighted all-cause hazard ratio compared to the effect of the Wei-Lachin approach. The Bakal approach lacks understandable interpretation. These results are supported by the application.

This general comparison and simulation study helps to understand the features of methods proposed for the analysis of composite endpoints combining (recurrent) events of different clinical relevance.

**AUTHORS/INSTITUTIONS:** A. Ozga, Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg- Eppendorf, Hamburg, GERMANY|H. Becher, Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, GERMANY|G. Rauch, Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Berlin, GERMANY|

**CONTROL ID:** 3385469

**TITLE:** Hierarchical Bayesian state-space modeling of age-and sex-structured wildlife population dynamics

**ABSTRACT BODY:**

**Abstract Body:** Modelling dynamics of wildlife or pastoral livestock population is important for developing a predictive understanding of their population change and hence for wildlife conservation and promoting human welfare. Developing reliable and realistic models for population dynamics of large herbivore population can be a very complex and challenging exercise. However, the Bayesian statistical domain offers some flexible computational methods that enable the development and efficient implementation of complex population dynamics models. In this work, we have used a novel Bayesian state-space model to analyse the dynamics of topi and hartebeest populations in the Serengeti-Mara Ecosystem of East Africa. The study is motivated by an age-sex structured population counts in different regions of Serengeti-Mara during the period 1989-2003. This analysis is aimed to detect the causes of recent declines in numbers of the herbivore populations that potentially threaten their future population viability in the ecosystem.

**AUTHORS/INSTITUTIONS:** S. Mukhopadhyay, Quantitative methods, Indian Institute of Management, Udaipur, Rajasthan, INDIA|H. Piepho, J. Ogotu, Biostatistics, University of Hohenheim, Stuttgart, Baden Wurtemberg, GERMANY|

**CONTROL ID:** 3385505

**TITLE:** Estimating Logistic Regression Parameters for Complex Survey Data: a Comparative Study.

**ABSTRACT BODY:**

**Abstract Body:** Complex survey data are becoming relevant in a number of fields. In data collected based on a complex sampling design, each observation has assigned a sampling weight, which is the number of subjects that this observation represents in the population. These data are commonly used for modelling purposes as data derived from a simple random sampling, but the effect of the sampling weights in the modelling process should be studied carefully. For instance, in the context of logistic regression models for dichotomous response variables, the pseudo-likelihood function (Binder, 1983) has been proposed as a modified version of the likelihood function that incorporates the sampling weights in the estimation process of the model parameters. Let us denote these models as the weighted logistic regression models. However, it is not yet clear if this method outperforms other alternatives when estimating the model parameters to complex survey data.

A simulation study has been conducted in order to compare the performance of different methods when estimating the model coefficients for a dichotomous response variable. Two pseudo-populations, with the covariates and response variables, have been generated based on two real surveys. The pseudo-populations have been sampled  $r=500$  times by one-step stratified sampling, replicating the sampling designs of the real surveys. One of these designs is informative (i.e., the selection probabilities of the subjects are related to the response variable of the survey), while the other is non-informative. For a given set of covariates, three models have been fitted to each sample: a) the ordinary logistic regression model, b) the generalized linear mixed model with random intercept for each sampling stratum, and c) the weighted logistic regression model. The parameters of these models have been compared to the true coefficients, which are the ones obtained by fitting the ordinary logistic regression model to the corresponding pseudo-population.

Summing up the conclusions, for informative design, ignoring the sampling weights in the likelihood function results in biased parameters with a large MSE. In contrast, for non-informative design, the results obtained by the three abovementioned methods are very similar in terms of bias and MSE. Therefore, the results suggest the use of the weighted logistic regression model.

**AUTHORS/INSTITUTIONS:** A. IPARRAGIRRE, I. BARRIO, I. AROSTEGUI, Applied mathematics, statistics and operations research, University of the Basque Country UPV/EHU, Leioa, SPAIN|I. AROSTEGUI, BCAM - Basque Center for Applied Mathematics, Bilbao, SPAIN|

**CONTROL ID:** 3385570

**TITLE:** Optimal design of survival multi-state models under constraints through finite state automata

**ABSTRACT BODY:**

**Abstract Body:** When considering complex pathologies like neuro-degenerative diseases (e.g. Alzheimer) or cancer-related genetic syndromes (e.g. Lynch syndrome), multi-state survival models are a natural way to take into account the clinical progression of the pathology (e.g. mild-cognitive impairment, dementia) or the spectrum of related diseases (e.g. neo-cancer locations).

In this context, it is often necessary to add constraints to the transitions between states. For example, a particular clinical state might occur only after a specific previous clinical evolution, the number of occurrence of certain type of recurrent events might be limited, or certain cancer localisation might be impossible after a specific event, and transitions toward a specific state might depend on the presence/absence of a specific event in the clinical past of the patient.

When the number of states and constraints is small enough, practitioners usually design the final multi-state model in an ad-hoc manner, but the extension to complex pathologies with a large number of such states and constraints is non trivial. We present here a general method allowing to build such multi-state model optimally and automatically from a set of constraints expressed as regular expressions. These expressions encode both the constraints of the model and the nature of the transitions.

For example, if we denote by 0 the healthy state, by 1 the death state, and by A and B two diseases, we can easily integrate the following rules/constraints: no more than two occurrences of A, no more than one occurrence of B, specific hazard rate of the second occurrence of A, specific hazard rate for the death depending of the previous occurrence of B.

Inspired from the finite state machine theory, these regular expressions are first turned into Nondeterministic Finite Automata (NFA) which are combined to build a Deterministic Finite Automata (DFA) representing the full multi-state survival model. A final minimization step is performed by identifying the equivalent-class of states. With the toy-example described above, a total of 8 NFAs allow to build a DFA with 7 states after minimization.

The interest of the method is illustrated in the context of the Lynch syndrome, a germline mutation of the one of the four mismatch repair genes which cause microsatellite instability cancers of a large spectrum: colorectal, gastric, endometrial, urothelial, etc.

**AUTHORS/INSTITUTIONS:** G. Nuel, A. Lefebvre, LPSM (CNRS 8001), Sorbonne University, Paris, FRANCE|O. Bouaziz, MAP5 (CNRS 8145), University of Paris, Paris, FRANCE|

CONTROL ID: 3385665

**TITLE:** Estimating the number of non-exposed cases using the population disease rate – limitations, data application, and comparison of methods

**ABSTRACT BODY:**

**Abstract Body:** Mortality statistics in industrialized countries are a sufficiently reliable and extensive data source for epidemiology and public health. For a specific year they provide information about frequencies of causes of death, usually also by categories of age at death and sex. However, if any other risk factor for the death due to a specific disease shall be investigated, this information is commonly not recorded and therefore the use of this data for epidemiologic research is limited.

We assume sufficient information on the population prevalence estimates  $p_{0jk}, p_{1jk}, p_{2jk}, \dots, p_{Mjk}$  for an exposure  $X$  with  $M+1$  categories  $0, 1, \dots, M$  for sex  $j$  and age group  $k$ , just as on the population figures  $n_{jk}$ . In some settings, additionally an estimate of the population rate of a specific disease is sufficiently known, for example for lung cancer where several studies and meta-analyses exist. Given such age and sex specific disease rates in the non-exposed  $\lambda_{0jk}$ , we can derive an estimate of the total number of non-exposed cases  $d'_{0..}$  as:  $d'_{0..} = \sum_{j,k} d'_{0jk} = \sum_{j,k} \lambda_{0jk} n_{jk} p_{0jk}$ , where the population figures  $n_{jk}$  serve as a proxy for the non-diseased population given the rare disease assumption.

As this approach has the inherent limitation of relying on the potentially biased or imprecise disease rate estimate among the non-exposed, we additionally develop a sensitivity analysis based on further data sources as follows. Given an estimate of the dose-response relationship between the specific exposure and the disease, just as the exposure prevalence estimates, we can estimate the total number of expected cases  $d'$  as:  $d' = \sum_m d'_{m..}$ , where  $d'_{m..} = \sum_{j,k} d'_{mjk} = \sum_{j,k} \lambda_{0jk} OR_m n_{jk} p_{mjk}$  and  $OR_m$  is the relative effect (risk or odds ratio) of exposure category  $m$  compared to the non-exposed category  $0$  and  $OR_0 = 1$ . The ratio between the observed total number of cases  $d$  and the total estimated number of cases  $d'$  can then be used to calibrate the disease rate estimate, and therefore to better estimate the number of non-exposed cases.

The method and its calibration are applied to the estimation of the number of never smokers among lung cancer cases in Germany.

**AUTHORS/INSTITUTIONS:** A. Aigner, H. Becher, Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, Berlin, GERMANY|H. Becher, Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, GERMANY|

**CONTROL ID:** 3385699

**TITLE:** Neural networks and regularization - with application to a three-generation study of BMI

**ABSTRACT BODY:**

**Abstract Body:** Introduction.

One of the most common types of deep neural networks is the multilayer perceptron (MLP) or feed forward neural networks. The architecture of a neural network refers to the number of layers and number of neurons in each layer. The main aim of regularization in a neural network (NN) is to prevent over-fitting. There are different methods for regularization. These include early stopping, the use of weighted l1 and l2 norms on the weights in the loss function, dropout and data augmentation. Questions that arise are:

1. does a NN with a complicated architecture and no regularization achieve the same accuracy as a NN with a simpler architecture and regularization?
2. to what extent are the regularization methods equivalent?

**Methods**

The different regularization methods are compared in a simulation study but more importantly NN's with and without regularization are compared. It maybe the most important hyper-parameters in a NN are the choice of number of hidden layers and neurons and the choice of activation function, with regularization of secondary importance. The models are applied to the unique three-generation family Lifeways study to predict BMI of children.

The h2o R package provides access to the h2o distributed Java-based implementation of a multilayer perceptron with many advanced features. We shall use h2o with MLP for the analyses as well as the keras R package.

**Results.**

Theoretical results regarding regularization obtained by Bishop (1995) are confirmed and augmented. Results regarding NN architecture and varying the activation function over the neurons in a layer are also obtained and compared to regularization results. Using data from the Lifeways cross-generation study, the body mass index (BMI) of children is predicted from that of their parents and maternal and paternal grandparents, via NN's with different hyper-parameters, linear models and generalized linear models. NN's give the best predictions of the methods considered.

**Reference.**

C. M. Bishop (1995). Regularization and complexity control in feedforward networks, Aston Univ., Tech. Rep. NCRG/95/022.

**AUTHORS/INSTITUTIONS:** G.E. Kelly, School of Mathematics and Statistics, University College Dublin, Dublin, IRELAND|

**CONTROL ID:** 3385712

**TITLE:** Self-organising maps in biology

**ABSTRACT BODY:**

**Abstract Body:** Self-organising maps have attractive properties for exploratory analysis of complex data, including intuitive visualizations and flexible grouping options. The implementation in the kohonen package for R [1] provides a full repertoire of standard SOM methodology but has additional functionalities that are potentially very useful in this Big Data era. If different sets of variables are present for the same objects, these can be arranged in different layers, where distance measures can be defined for each layer individually. Layers can also be given different weights (including zero) allowing for quick and intuitive assessments of layer correspondence.

The presentation will highlight the essentials of the kohonen package and show illustrative case studies in plant sciences.

[1] "Flexible Self-Organizing Maps in kohonen 3.0", R. Wehrens and J. Kruisselbrink, J. Stat. Softw. 87, 7 (2018)

**AUTHORS/INSTITUTIONS:** R. Wehrens, Biometris, Wageningen UR, Wageningen, NETHERLANDS]

**CONTROL ID:** 3385723

**TITLE:** A regression analysis between food preferences and breast cancer deaths in public applying the age-environment model to age-by-period data

**ABSTRACT BODY:**

**Abstract Body:** Food preferences are important factors in personal health, implying an instrumental relationship between food and health. Especially for breast cancer eating habit is one important factor to lower or higher your risk of the disease. Actually, the American Cancer Society recommends eating mostly vegetables, fruits, and whole grains, and less red meat, and fewer sweets. In our previous study, also for public health, the trend of breast cancer deaths seemed to have a correlation with the trend of preference rates for beef stake in public in Japan. In this study, considering those facts, we propose a method for regression analysis between food preferences and breast cancer deaths in public applying the age-environment model to age-by-period data for food preferences and breast cancer death rate, and apply the proposed method to actual data obtained in Japan. The result suggests that that the change of food preference from beef stew to Hamburger and beef stake increase the breast cancer risks in Japan.

**AUTHORS/INSTITUTIONS:** N. Hanayama, Information Expression, Shobi University, Kawagoe, Saitama, JAPAN|

**CONTROL ID:** 3385806

**TITLE:** Real-world scenarios in multivariate analysis: the impact of sample size and imbalance on power in silico

**ABSTRACT BODY:**

**Abstract Body:** Genome-wide association studies uncovered many disease-associated loci through univariate model, with the focus on single trait. As the expansion of biobank cohorts and increasing availability of the wide spectrum of disease status, multivariate methods have recently been proposed to effectively investigate the relationships between a genetic variant and multiple diseases. These methods assist in the discovery of an important biological phenomenon – pleiotropy, which occurs when a variant influences multiple traits. Most of the evaluation of multivariate methods have been focused on continuous phenotypes or balanced case-control phenotypes. However, the statistical performance on unbalanced case-control sample size, which is always seen in natural data such as electronic health records, is largely unknown. In this work, we designed a large-scale, real-world informed multivariate simulation study for binary phenotypes to thoroughly characterize the type I error and power for multivariate methods. Meanwhile, we compared their performance with the standard univariate method for all simulation settings. The simulation parameters were set according to the summary statistics obtained from UK Biobank, including minor allele frequencies, genetic effect sizes, disease prevalence and phenotypic correlation structure. Our results demonstrate the statistical characteristics of univariate and multivariate methods under a wide range of sample size and imbalance design, and can serve as a reference guide to the application of multivariate analysis in natural dataset.

**AUTHORS/INSTITUTIONS:** X. Zhang, R. Li, M. Ritchie, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES|X. Zhang, M. Ritchie, Genetics, University of Pennsylvania , Philadelphia, Pennsylvania, UNITED STATES|X. Zhang, Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES|

**CONTROL ID:** 3385807

**TITLE:** A correct way of handling mixture model-driven data for causal inference

**ABSTRACT BODY:**

**Abstract Body:** As data science advances, reducing high-dimensional data into meaningfully parsimonious classes is a principal task for studying associations among a target response variable and its associated high-dimensional causes. The reduction of high-dimensional data has been readily done by parametric mixture models such as Latent Class Analysis (LCA) and Latent Dirichlet Allocation (LDA) models. However, statistical inference with mixture model-driven variables needs prudence because they assign probabilities of cluster memberships, which are not directly observable and hence are missing data. Thus, a missing data strategy is needed to handle mixture model-driven clusters correctly. This presentation will discuss a statistical framework called expected estimating equation (Kang et al., 2019) to draw a proper statistical inference with mixture model-driven clusters. Examples are estimating the causal effects of text data with LDA and estimating causal effects for an infectious disease using LCA. LDA was applied to survey text notes with survey response status in the Survey of Income and Program Participation (SIPP). LCA was applied to the National Health and Nutrition Examination Survey (NHANES).

**AUTHORS/INSTITUTIONS:** J. Kang, Center for Optimization and Data Science, U.S. Census Bureau, Suitland, Maryland, UNITED STATES|J.L. Schafer, ADRM, U.S. Census Bureau, Suitland, Maryland, UNITED STATES|

**CONTROL ID:** 3385878

**TITLE:** Non-linear feature detection in biometric applications

**ABSTRACT BODY:**

**Abstract Body:** Functional magnetic resonance imaging (fMRI) allows for the effective measurement of the whole brain activity. Correct assessment of the autocorrelations between the regions in the brain, i. e. functional connectivity, might be useful in understanding how a person performs a certain task. However, testing the significance of changes in functional connectivity is a challenging task since fMRI series exhibit both temporal and spatial dependence. To address this issue we propose a bootstrap procedure based on non-linear dimension reduction in frequency domain. Our choice of non-linear dimension reduction is justified by a well-known fact that the brain is a non-linear system. Another well-known feature of fMRI series is long-range dependence, therefore using discrete wavelet-transform we convert long-range dependence in the time-domain to the short-range dependence in the wavelet domain. Finally, we propose a new statistical bootstrap hypothesis testing framework after taking non-linearity and long-range dependence of the data into account. The performance of the suggested procedures are tested in simulations and on real data.

**AUTHORS/INSTITUTIONS:** N. Sirotko-Sibirskaya, University of Bremen, Bremen, Bremen, GERMANY|

**CONTROL ID:** 3385924

**TITLE:** Classification of human physical activity based on the raw accelerometry data via spherical coordinate transformation

**ABSTRACT BODY:**

**Abstract Body:** Human health is strongly associated with person's lifestyle and levels of physical activity. Therefore, characterization of daily human activity is an important task. Accelerometers have been used to obtain precise measurements of body acceleration. Wearable accelerometers collect data as a three-dimensional time series with frequencies up to 100Hz. Using such accelerometry signal, we are able to classify different types of physical activity. In our work, we present a novel procedure for physical activity classification based on the raw accelerometry signal. Our proposal is based on the spherical representation of the data. We classify four activity types: resting, upper body activities (sitting), upper body activities (standing) and lower body activities. The classifier is constructed using decision trees with extracted features consisting of spherical coordinates summary statistics, moving averages of the radius and the angles, radius variance and spherical variance.

The classification accuracy of our method has been tested on data collected on a sample of 47 elderly individuals who performed a series of activities in laboratory settings. The achieved classification accuracy is over 90% when the subject-specific data are used and 84% when the group data are used. Main contributor to the classification accuracy is the angular part of the collected signal, especially spherical variance. To the best of our knowledge, spherical variance has never been previously used in the analysis of the raw accelerometry data. Its major advantage over other angular measures is its invariance to the accelerometer location shifts.

**AUTHORS/INSTITUTIONS:** J. Harezlak, Epidemiology and Biostatistics, Indiana University, Bloomington, Indiana, UNITED STATES|J. Harezlak, M. Kos, M. Bogdan, Mathematics, University of Wroclaw, Wroclaw, POLAND|N.W. Glynn, Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania, UNITED STATES|

**CONTROL ID:** 3385969

**TITLE:** Juror understanding of statistical evidence: a systematic review.

**ABSTRACT BODY:**

**Abstract Body:** The interpretation of statistical methods as it relates to forensic evidence plays a significant role in the smooth and just operation of the criminal justice systems that adopt common law. There has been a recent increase in inquiries into the applied methods used to report statistical evidence and the way such evidence is presented to a lay audience. However, there is neither consensus on the most appropriate approach of presenting evidence in practice, nor an in-depth view of how laypeople understand probability and statistics in varying contexts.

Numerous attempts have been made to understand how jurors interpret statistical evidence that offer different opinions on the statistical analysis of forensic evidence and the ability of laypeople to successfully comprehend statistical evidence. The results of a systematic literature review will be presented and aimed to critically evaluate those attempts and to provide an in-depth view of existing knowledge on the topic. Two streams of acquisition of articles for inclusion are identified: database searching, and articles already known to the research team. Articles already known to the research team will be presented as preliminary results in this paper, with final results to be presented at IBC2020.

This article will contribute knowledge to the literature on how jurors understand and interpret statistical evidence. The significance of this research will be the provision of knowledge to establish a baseline understanding for the ultimate goal of assisting all those involved to improve the way statistical evidence is presented for optimal understanding by jurors.

**AUTHORS/INSTITUTIONS:** A.V. Cronin, J. Williams, N. Subramaniam, D. Vagenas, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, AUSTRALIA|B. Matthews, School of Law, Queensland University of Technology, Brisbane, Queensland, AUSTRALIA|J. Williams, School of Environment and Science, Griffith University, Brisbane, Queensland, AUSTRALIA|

**CONTROL ID:** 3386007

**TITLE:** Interpretation of autoregressive linear mixed effects models with time as a continuous covariate

**ABSTRACT BODY:**

**Abstract Body:** In an autoregressive linear mixed effects model, the current response is regressed on the previous response, fixed effects, and random effects. We previously proposed this model and showed the interpretation (Funatogawa et al., 2007; Funatogawa and Funatogawa, 2018). The model has two remarkable properties: approaching an asymptote and state dependency. An autoregressive linear mixed effects model with a random effect in initial measurement and a random intercept corresponds to a monomolecular growth curve with a random baseline and a random asymptote. It shows a nonlinear time trend with an initial rapid change. The asymptote is expressed as fixed and random effects divided by 1 minus an autoregressive coefficient. When the values of explanatory variables are changed, the response moves to a new asymptote. However, when the model includes time as a continuous explanatory variable, the interpretation of the model is unclear. In this study, we present the interpretation and characteristics of the autoregressive linear mixed effects model with time as a covariate. The response trajectory is expressed by a sum of a monomolecular growth curve and a linear time trend. The asymptote mentioned above is more appropriately referred to as target. The target is a linear time trend. The response moves to the target values, but never catch up. After a sufficient time, the response is expressed by a linear time trend which is parallel to the linear time trend of the target values. This model can well represent data that have a non-linear initial change and a linear change after a sufficient time. The random effects can represent a random variation in baseline and random variations in the intercept and slope for the linear time trend. Because longitudinal data have often a limited number of measurement points, the curve at the early phase before reaching steady state is important compared with time series data with a large number of measurement points. The clear image of the trajectories that the model can represent will help appropriate modeling.

**AUTHORS/INSTITUTIONS:** I. Funatogawa, Department of Data Science, The Institute of Statistical Mathematics, Tokyo, JAPAN|T. Funatogawa, Clinical Science and Strategy Department, Chugai Pharmaceutical Co., Ltd., Tokyo, JAPAN|

**CONTROL ID:** 3386077

**TITLE:** Using Automated Machine Learning for Predicting Individual Risk of Malignancy in the Patients with Intraductal Papillary Mucinous Neoplasms

**ABSTRACT BODY:**

**Abstract Body:** Intraductal papillary mucinous neoplasms (IPMN) are premalignant lesions of the pancreas. Although clinical guidelines were released in 2012 to improve diagnosis, treatment for IPMN, due to vague terminology and insufficient data, classifying IPMN and assigning individual risks of malignancy to each patient remain unclear. To evaluate an individual risk of malignancy and to classify IPMN into benign or malignant groups, we used large database of 3,464 patients from 31 different hospitals. This study was a multi-national (Korea, Japan, United States, China, Sweden, and Taiwan) retrospective study. We then used automated machine learning to choose the best machine learning algorithm for classifying IPMN patients. Seven algorithms of machine learning (XG boost, deep learning, distributed random forest, extremely randomized trees, generalized linear model, gradient boosting machine, and stacked ensemble) were utilized and compared. This research systematically analyzed which machine learning algorithm to choose for classifying IPMN using automated machine learning.

**AUTHORS/INSTITUTIONS:** C. Lee, Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|T. Park, Department of Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|J. Kang, J. Jang, Department of Surgery and Cancer Research Institute, Seoul National University College of Medicine, Seoul, KOREA (THE DEMOCRATIC PEOPLE'S REPUBLIC OF)|

**CONTROL ID:** 3386358

**TITLE:** Are random coefficient models useful for predicting cultivar performance in an untested site?

**ABSTRACT BODY:**

**Abstract Body:** Multi-environment trials (MET) are conducted to assess the performance of a set of cultivars in a target population of environments (TPE). From a grower's perspective, MET results must provide predictions of cultivar performance in untested sites, i.e. his/her sites, and these do not coincide with the sites at which the trials were conducted. Linear mixed modelling can provide predictions for untested sites. However, the precision of the predictions is of primary concern and should be assessed. The precision can be improved when more information is given to characterize the targeted sites. Thus, in this study, we demonstrate the benefit of using environmental information (covariates) for predicting cultivar performance in an untested site for Swedish winter wheat official trials. Furthermore, Swedish MET sites can be stratified into zones, which allows borrowing information between zones when best linear unbiased prediction (BLUP) is used. To account for correlations between zone intercepts and slopes for the covariates, we fitted random coefficient models. The results showed that by using covariates and borrowing information between zones improved the precision of predictions for untested sites. To our knowledge, this study is the first using random coefficient for predicting cultivar performance in untested sites based on MET.

**AUTHORS/INSTITUTIONS:** H. Buntaran, H. Piepho, Biostatistics, Institute of Crop Science, University of Hohenheim, Stuttgart, Baden-Württemberg, GERMANY|J. Forkman, Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Uppsala, SWEDEN|

**CONTROL ID:** 3386392

**TITLE:** A Robust Predictive Multivariate Monitoring using the Preliminary Limits and the Robust Principal Component Analysis Biplot: A New Combined Algorithm.

**ABSTRACT BODY:**

**Abstract Body:** A peculiar method of the Biplot approximation is the principal component analysis (PCA) Biplot that preserves both the PCA properties and the multidimensional representation of the objects (samples) and their corresponding axes (variables) on a single plot. In order to develop a modern predictable and robust multi-purpose online multivariate process monitoring chart, a new combined proposal of the preliminary limits and the robust PCA Biplot is exploited. With  $L: 2 \times p$  preliminary matrix comprising the upper and lower limits for  $p$  process variables, an integrated algorithm that superimposes the robust PCA Biplot of the new process dataset on the grid is devised using noteworthy robust singular value decomposition PCA approaches. The resulting configuration, which is constrained by the  $L$  matrix, becomes the cornerstone for a user defined predictive multivariate monitoring regions with  $p(p-1)+2$  total regions on which guided predictions could be made from. The new method is appraised by both simulation studies and empirical applications from tobacco manufacturing process datasets, and results obtained revealed promising schemes that fostered quality decision making.

**AUTHORS/INSTITUTIONS:** C. John, E.J. Ekpenyong, C.O. Omekara, Statistics, Michael Okpara University of Agriculture, Umudike, Umuahia, Abia State, NIGERIA|

**CONTROL ID:** 3386543

**TITLE:** Evolutionary model of protein domains in bacterial genomes

**ABSTRACT BODY:**

**Abstract Body:** Protein domains are components of protein that can fold independently into a compact and stable structure and are considered the fundamental units of proteins. As pointed out by Xie et al. (2011), evolutionary changes of protein domains are some of the main mechanisms that allow the emergence of proteins with new functionalities. For this reason, the investigation of the dynamic processes that led to the current configuration of protein domains can highlight the important aspects of the proteome evolution and consequently of the evolution of living organisms.

The genome sequences of ~3000 bacteria were downloaded from the NCBI database and protein domain identification was performed for every sequence resulting in the list of protein domain frequencies for each bacteria. Three different evolutionary hypotheses were tested by fitting the Relative Species Abundance (RSA) distribution of the protein domain families with three different distributions: the Poisson Log-Normal, the Negative Binomial and the Log-Series. Based on the Akaike Information Criterion, the protein domain RSA is best described by Poisson Log-Normal model, firstly introduced by Engen and Lande in 1996, which assumes an underlying birth-death process subject to both additive (demographic) and multiplicative (environmental) noise.

To assess the ability of the RSA model in describing the proteome evolution, the hierarchical clustering based on model parameters,  $\mu$  and  $\sigma$ , and density of protein domains was performed. Comparison of results with bacterial taxonomy and 16S rRNA based phylogeny showed a good agreement both with taxonomy and phylogeny, based on corrected Normalized Mutual Information. However, the resolution at which the RSA model gave the most interesting results is at the species level, identifying groups of divergent bacterial strains. For example, the RSA model was able to distinguish two different pathotypes among 21 strains of *Xanthomonas citri*. Although microbiological taxonomy is mostly based on the hypervariable region of the 16S rRNA gene which serves as a good estimate of evolutionary time, bacterial classification based on 16S rRNA gene has low phylogenetic power at species level where functional diversification of strains is faster than random mutations of 16S rRNA gene. Since the proposed RSA model is estimating the proteome of bacteria species, it manages to mitigate these shortcomings.

**AUTHORS/INSTITUTIONS:** I. Budimir, C. Sala, D. Remondini, Department of Physics and Astronomy, University of Bologna, Bologna, ITALY|E. Saccenti, M. Suarez-Diez, Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, NETHERLANDS|G. Castellani, Department of Experimental, Diagnostic and Specialty Medicine - DIMES, University of Bologna, Bologna, ITALY|

**CONTROL ID:** 3386578

**TITLE:** A WEIBULL-RAYLEIGH BASED GINI COEFFICIENT TO MONITOR ADHERENCE TO MALARIA PREVENTION AND TREATMENT IN NIGERIA

**ABSTRACT BODY:**

**Abstract Body:** The study derived the Gini index of the Weibull – Rayleigh distribution and apply the coefficient to monitor the critical indicators from the Nigerian Malaria Indicators Survey (NMIS). We explore the pattern of inequality in adherence to malaria prevention and treatment strategies for children using relevant indicators of malaria prevention and treatment as extracted from the 2015 NMIS across the six geo-political regions of Nigeria. Gini coefficients were obtained using both non-parametric and parametric approaches. The Weibull-Rayleigh distribution based Gini coefficient used for the parametric approach was applied to the scaled data. Results show uniformity of pattern of adherence across the nation with Gini index of less than 10%. This implies that adherence pattern to malaria prevention and treatment for children is almost the same across and within all the geo-political regions in Nigeria. Plots of the Lorenz curve also confirm the results. The findings based on this research work will provide a new direction for policy makers in Nigeria on meeting the National Strategic Plan for Malaria Control targets.

**AUTHORS/INSTITUTIONS:** E.E. AKARAWAK, Mathematics, University of Lagos, Lagos, NIGERIA|I.A. ADELEKE, Actuarial Science & Insurance, University of Lagos, Lagos, NIGERIA|

**CONTROL ID:** 3386581

**TITLE:** SCISSOR: a novel framework for identifying structural changes in RNA transcripts

**ABSTRACT BODY:**

**Abstract Body:** We propose a statistical method, SCISSOR, for unsupervised screening of a range of structural alterations in RNA-seq data. Compared to other existing methods relying on a limited subset of RNA-seq data available, e.g. exon/gene level expression or junction split reads, we consider a novel shape property of aligned short read data through a base-level pileup file. This intact and uncompressed view of RNA-seq profile enables the unbiased discovery of structural alterations by looking for anomalous shapes in expression.

Shape changes in selecting sample outliers in RNA-seq, SCISSOR, is a series of procedures for transforming and normalizing base level RNA sequencing coverage data in a transcript independent manner, followed by a statistical framework for its analysis. The resulting high dimensional object is amenable to unbiased identification of structural alterations across RNA-seq cohorts with nearly no assumption on the mutational mechanisms underlying abnormalities. This enables SCISSOR to independently recapture known variants such as splice site mutations in tumor suppressor genes as well as novel variants that are previously unrecognized or difficult to identify by any existing methods including recurrent alternate transcription start sites and recurrent complex deletions in 3' UTRs.

In a cohort of 522 TCGA head and neck squamous cell carcinoma, SCISSOR identifies known as well as novel aberrations including abnormal splicing, intra-/intergenic deletions, small indels, and alternative transcription start/termination. In addition, its genome-wide analysis uncovers a novel type of variant gene transcription near intragenic CpG islands. Finally, we find that our approach through shapes can be also useful for picking up rare transcripts from individual samples such as leukocyte transcripts and stromal transcripts. We believe that this new approach holds promise for identifying otherwise obscured genetic aberrations.

Taken together, these results suggest that SCISSOR has great potential for broad applications including discovery of novel driver genes and mechanisms of genetic abnormalities, detection of non-coding RNA, and studies of single cell RNA-seq.

**AUTHORS/INSTITUTIONS:** H. Choi, D.N. Hayes, UTHSC, Memphis, Tennessee, UNITED STATES|H. Choi, D.N. Hayes, Computational Biology, St. Jude Children's Research Hospital, Memphis, Tennessee, UNITED STATES|J.S. Marron, Statistics and Operations Research, UNC, Chapel Hill, North Carolina, UNITED STATES|

**CONTROL ID:** 3386622

**TITLE:** CLUSTER VALIDATION TO IDENTIFY POPULATION GENETIC STRUCTURE FROM MASSIVE GENOMIC DATA

**ABSTRACT BODY:**

**Abstract Body:** Current technologies in genomics have enabled to generate large volumes of data that have thousands of variables characterizing a biological unit. Identifying population genetic structure (PGS) from genomic data is crucial for breeding and conservation purposes. Several clustering algorithms are available for use with genomic data to group several genotypes and depict PGS. In this work, we compared the performance of three clustering methods –Unweighted Pair Group Method with Arithmetic Mean (UPGMA), k-means and Structure Bayesian method– and four validation indices –connectivity, Dunn, Ch, and silhouette–to detect PGS and the reliable number of groups defining that PGS. We simulated a dataset to know the real number of groups of a population and the group each individual was assigned to. Molecular databases from Single Nucleotide Polymorphism (SNP) markers for diploid individuals were simulated using "Xbreed" package of R. Three PGS scenarios were generated, differing in the number of subpopulations:  $k=2$ ,  $k=5$  and  $k=10$ , with  $F_{st}=0.03$  as level of divergence between them. For each scenario, 100 replicates of 1000 individuals and 80K SNPs each were simulated. We evaluated the percentage of wrong classification for each scenario and for each clustering method. We implemented the clustering methods for  $k=2$  to  $k=15$  to evaluate the underestimation and overestimation of the group number error (Error Type III) for each validation index. Structure Bayesian method was the best to classify individuals in all scenarios, whereas UPGMA was the worst. In those scenarios with  $k=5$  and  $k=10$ , connectivity had the maximum underestimation of group number error for all cluster algorithms. The indices with best performance to validate number of groups were Silhouette and Dunn in all scenarios using the Bayesian method.

**AUTHORS/INSTITUTIONS:** M.E. Videla, Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Córdoba, Córdoba, ARGENTINA|M.E. Videla, Cátedra de Estadística y Biometría. Facultad de Ciencias Agropecuarias (FCA). Universidad Nacional de Córdoba. Group associated with UFyMA (INTA -CONICET)., Córdoba, ARGENTINA|M.E. Videla, Universidad Nacional de Villa María, Argentina, Villa María - Córdoba, ARGENTINA|C.I. Bruno, Biometric and Statistics, National University of Córdoba, Córdoba, Córdoba, ARGENTINA|

**CONTROL ID:** 3386624

**TITLE:** Spatiotemporal kriging to interpolate the precipitation values in Paraíba, Brazil

**ABSTRACT BODY:**

**Abstract Body:** The north-eastern region of Brazil (NEB) is characterized by an irregular, highly variable distribution of rainfall in space and time. In this region, it is common to find high rates of rainfall at locations adjacent to those with no record of rain. This paper investigates the spatiotemporal distribution of precipitation within NEB, in the state of Paraíba, Brazil. Paraíba experiences localized periods of drought within rainy seasons and distinct precipitation patterns among the state's mesoregions. The datasets used in this study were obtained from the Executive Agency of Water Management (AESAs) of the State of Paraíba, Brazil, which is responsible for rainfall information in the region. The state of Paraíba spans an area of approximately 56,585 km<sup>2</sup> in north-eastern Brazil between the 6° and 8° parallels of south latitude and the 34° and 39° meridians of west longitude. The state of Paraíba is situated in a tropical region, and it is divided into the following four mesoregions: Zona da Mata, Agreste, Borborema and Sertão. The dataset used in the present study includes the time series of the total monthly rainfall recorded at 269 rain gauge stations from 1994 to 2014. The mean precipitation values observed at several irregularly spaced rain gauge stations from 1994 to 2014 showed remarkable variations among the mesoregions in Paraíba throughout the year. As a consequence of this behavior, there is a need to model the rainfall distribution jointly with space and time. A spatiotemporal geostatistical methodology was applied to monthly total rainfall data from the state of Paraíba. The rainfall data indicate intense spatial and temporal variabilities that directly affect the water resources of the entire region. The results provide a detailed spatial analysis of sectors experiencing precipitation conditions ranging from a scarcity to an excess of rainfall. The present study should help drive future research into spatiotemporal rainfall patterns across all of NEB.

**AUTHORS/INSTITUTIONS:** R.R. de Lima, Department of Statistics, Federal University of Lavras, Lavras, Minas Gerais, BRAZIL|E.S. de Medeiros, FACET, Federal University of Grande Dourados, Dourados, MS, BRAZIL|R.A. de Olinda, Department of Statistics, Paraíba State University, Campina Grande, PB, BRAZIL|C.C. dos Santos, Center for Technology and Natural Resources, Federal University of Campina Grande, Campina Grande, PB, BRAZIL|C.C. dos Santos, Daugherty Water for Food Global Institute, University of Nebraska, Lincoln, Nebraska, UNITED STATES|

**CONTROL ID:** 3386636

**TITLE:** Homogeneity Analysis of In Situ Precipitation Data for Climate Services in Kenya

**ABSTRACT BODY:**

**Abstract Body:** The historical climatic data provide a baseline for studies of climate change. To provide this baseline they must be of high quality and an important aspect of the quality is that they be homogeneous. Climate records need to be prepared well to remove any artificial inhomogeneities which hide the real changes of climate and may lead to wrong conclusions. Long-term climatological series often contain non-climatic jumps. Kenyan meteorological data is not an exception as over the years there have been changes in exposure of measuring instruments, the introduction of automatic weather stations, urbanization and land-use changes among other factors. The homogeneity of monthly precipitation data from meteorological stations in Kenya was assessed using Craddock test, Pettit test, Buishand range test, Von Neumann's ratio test, Standard Normal Homogeneity Test (SNHT) and Bayesian test. SNHT is a widely used technique and this study confirmed it is an effective method. Breakpoints were detected from the long term precipitation series which were adjusted and corrections applied to produce homogeneous climatological series.

**AUTHORS/INSTITUTIONS:** S.K. Magut, Statistics, AIMS- Cameroon, Kisumu, KENYA|S.K. Magut, Programming, African Maths Initiative, Kisumu, Nyanza, KENYA|

**CONTROL ID:** 3386641

**TITLE:** Confidence intervals for the difference of two independent folded normal distributions for the comparison: derivation and application

**ABSTRACT BODY:**

**Abstract Body:** The absolute change in the corrected angle measured immediately after surgery and after bone healing is a clinically relevant endpoint to judge stability of an osteotomy. To demonstrate non-inferiority of a novel screw used for fixation of the osteotomy in comparison with a standard screw, may be the main aim of a clinical study. If the difference in the angles after surgery and after bone healing can be assumed to be normally distributed, the absolute change follows a folded normal distribution. The most natural approach for demonstrating non-inferiority would be the use of a confidence interval for the difference of two folded normal distributions. In this presentation we derive confidence intervals for the difference of two independent folded normal distributions using the delta method. We illustrate the approaches from a study on hallux valgus osteotomy. The proposed confidence intervals permit to investigate non-inferiority of two treatment groups in clinical trials with endpoints following a folded normal distribution.

**AUTHORS/INSTITUTIONS:** A. Ziegler, Medizincampus Davos, Davos, SWITZERLAND|A. Ziegler, S. Ogutu, H.G. Mwambi, School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, SOUTH AFRICA|

**CONTROL ID:** 3386671

**TITLE:** Two step procedure to model site specific herbicide soil persistence

**ABSTRACT BODY:**

**Abstract Body:** Soil persistence is the length of time an herbicide remains active in soil and it is crucial to describe risks of diffuse contamination due to the application of herbicides to soil. Persistence is described by the parameter half-life, which is the time it takes to reach half of the initial concentration supplied to the soil. Its quantification demands the construction of dissipation curves from periodic determinations of the analyte on the same soil. This type of data is costly to obtain and has been usually modeled assuming the existence of a single variance component and independence of the observations used to adjust the curve, which makes it difficult to interpret them in a comprehensive environmental context. The objective of this work was to design a statistical workflow to explain environmental behavior of persistence. A sample of soils over a wide region under study was selected using the cLHS sampling method. The soils were fortified with atrazine and incubated for 21 days by measuring herbicide concentrations at days 0,3,7,14 and 21 by liquid chromatography coupled to tandem mass spectrometry (LC-MS / MS) using QuEChERs. A two-step procedure was performed to explain site specific half-life: First, estimation of the decay of the herbicide along time with a mixed nonlinear model with random effect of soil associated to the decay rate to obtain half-life from the decay curve for each soil sampled; and second a statistical modeling of the half-life estimated at each site based on soil properties and management using Bayesian spatial regressions (R-INLA). The addition of a random effect on the decay rate produced a better fit and provided a tool to explore half-life variability of between soils in a region which was further modeled by a Bayesian spatial regression. Atrazine soil persistence variability was mainly explained by agricultural land uses, where sites with crops that often use atrazine had higher decay rates. From this method integration it was possible to enhance the environmental understanding of herbicides persistence process.

**AUTHORS/INSTITUTIONS:** F. Giannini Kurina, M. Balzarini, Estadística y Biometría, Facultad de Cs. Agropecuarias, Universidad Nacional de Córdoba, Córdoba, Córdoba, ARGENTINA|F. Giannini Kurina, M. Balzarini, CONICET, Córdoba, Córdoba, Córdoba, ARGENTINA|J. Borello, CEPROCOR, Córdoba, Córdoba, ARGENTINA|S. Hang, National University of Córdoba, Córdoba, Córdoba, ARGENTINA|

**CONTROL ID:** 3386675

**TITLE:** Morphological delimitation of *Phyllanthus orbicularis* in Cuba using classic morphometry and geometric morfometry analysis of Fourier Elliptic Descriptors

**ABSTRACT BODY:**

**Abstract Body:** *Phyllanthus orbicularis* Kunth is an endemic plant in Cuba that is distributed throughout the country. From revisions of herbarium materials, morphological differences were found between the specimens of *P. orbicularis* from the western and eastern regions of the island and discrepancies were observed in the literature regarding the classification of the species. Due to the importance of this species in the treatment of some diseases and for conservation it is necessary to clarify its taxonomic state. That is why the present study aimed to verify whether there were morphological differences between the western and eastern variants of *P. orbicularis* that allowed their separation into independent taxa, from the measurement of morphological variables using classic and geometrical morfometry. Six locations were visited (Cajálbana, Galindo, Los Caneyes, Ceja de Melones, La Cueva and Yamanigüey) and samples were taken from branches, flowers and fruits of 10 individuals per location and 360 materials of 18 herbariums were reviewed too. Comparisons were made between the plants of the visited towns in terms of: number of leaves per floriferous branch, length of the floriferous branches, dimensions of the leaf blade, pedicel length, dimensions of the calyx pieces, stamens and styles length, ovary, fruits and seeds dimensions. Differences in leaf sheet contours between sampled plants from different locations were analyzed from the Principal Component Analysis of the coefficients of Fourier elliptical descriptors. The results obtained showed that there are significant differences between the plants of La Cueva and Yamanigüey, with respect to the plants of the rest of the visited locations, sufficient to be considered in a new species. The main characters that allowed the separation of this new species were the length of the floriferous branches, the number of leaves per floriferous branch, the shape of the leaf blade, flowers dimensions and pollen grain ornamentation. The characters of the new species were checked, as well as the characters defined for the *P. orbicularis* plants in the rest of the island and in correspondence with their original description in the herbarium materials consulted.

**AUTHORS/INSTITUTIONS:** D. de Vales Fernández, L. Pérez Pelea, Biología Vegetal, Facultad de Biología, Universidad de la Habana, La Habana, CUBA|B. Falcón Hidalgo, Docente-Investigativo, Jardín Botánico Nacional, Universidad de la Habana, La Habana, CUBA|

**CONTROL ID:** 3386678

**TITLE:** “But what about interactions, are any of those significant?”

Resolving the Collaborative Nightmares of Covariate Interactions through Regularization

**ABSTRACT BODY:**

**Abstract Body:** Sifting through all possible interactions in modeling applications is a dangerous statistical endeavor. All too often, one or more of the many interactions is found to “significantly” improve the fit, and we burden ourselves with trying to interpret an opaque model with interactions that do not make clinical sense. With this in mind, we explore and illustrate the concept of ranked sparsity, a phenomenon that often occurs naturally in the presence of derived variables such as interactions. In particular, ranked sparsity arises in modeling applications when an expected disparity exists in the quality of information between different feature sets. Its presence can cause traditional and modern model selection methods to fail because such procedures commonly presume that each potential parameter is equally worthy of entering into the final model – we call this presumption “covariate equipoise”. However, when all possible interactions are considered as candidate predictors, the premise of covariate equipoise will often produce over-specified and convoluted models. The sheer number of additional candidate variables grossly inflates the number of false discoveries in the interactions, resulting in unnecessarily complex and difficult-to-interpret models with many (truly spurious) interactions. We suggest a modeling strategy that requires a stronger level of evidence in order to allow certain variables (e.g. interactions) to be selected in the final model. This ranked sparsity paradigm can be implemented with the sparsity-ranked lasso (SRL). We compare the performance of SRL relative to competing methods for selecting interactions in a series of simulation studies, showing that the SRL is fast, accurate, and produces more transparent models (with fewer false interactions) than other state-of-the-art methods. We illustrate its utility in an application to predict the survival of lung cancer patients using a set of gene expression measurements and clinical covariates, searching in particular for gene-environment interactions, which are very difficult to find in practice.

**AUTHORS/INSTITUTIONS:** R.A. Peterson, Biostatistics & Informatics, University of Colorado School of Public Health, Denver, Colorado, UNITED STATES|J.E. Cavanaugh, Biostatistics, University of Iowa College of Public Health, Iowa City, Iowa, UNITED STATES|

**CONTROL ID:** 3386711

**TITLE:** Bayesian Multiple Change-point Estimation for Hazard in Weibull Distributions

**ABSTRACT BODY:**

**Abstract Body:** A Bayesian approach to the problem of hazard change with unknown multiple change-points is developed using informative and noninformative priors for survival data. Due to their flexibility, Weibull distributions are suitable for parametric modeling of time-to-event data and they have been widely used in scientific applications such as in survival analysis, reliability and industrial engineering, hydrology and etc.. For the Weibull distribution, piecewise constant hazard is considered with change-point estimation. The two-parameter Weibull distribution is a particularly difficult with related prior distribution because it requires a two-dimensional joint prior for Weibull parameters. The segment neighborhood (SN) algorithm is implemented for efficient search for change-points with the posterior distributions. A major feature of the proposed approach is that there is no necessity for the number of change-points resulting hazard function. The performance of the proposed estimator is checked via simulation. As a real data application, Leukemia data and Kidney data are analyzed.

**AUTHORS/INSTITUTIONS:** J. Kim, Statistics, Duksung Women's University, Seoul, KOREA (THE REPUBLIC OF)

**CONTROL ID:** 3386724

**TITLE:** Decision tree for influenza detection

**ABSTRACT BODY:**

**Abstract Body:** Background

Influenza is a contagious respiratory illness that seriously threatens public health as it has a high risk to become a pandemic. An active surveillance system that has a lower cost in money and time for accurate real-time prediction of influenza is required.

Objectives

We have shown that electronic swab form can be of value to improve influenza detection in real-time. This paper describes the use of a decision tree algorithm classifier for influenza detection and compares their diagnostic capabilities against an influenza PCR test.

Methods

We gather data from Montenegro primary and secondary care units in influenza season 2018 and 2019. Data is used from the form that is collected together with swab testing on influenza. We employed the decision tree algorithm to extract influenza-related findings on symptoms and to encode them into one of two values: influenza-positive and negative. The optimal prunings were done based on the complexity parameter value. The confusion matrix is used to evaluate the classification performance.

Results

The total number of observations was 1121 that we divided into two different datasets: training (60% cases), and test (40%). The final model was a simple tree with only four splits based on cough, temperature more than three days, temperature as a dichotomous variable ( $\geq 38^{\circ}\text{C}$ ) and sore throat. Accuracy with confusion matrix was 0.57, with 95% CI : (0.5267, 0.6217)

Our study demonstrates swab form in conjunction with the use of a decision tree can lower costs of influenza testing and might significantly reduce hospital expenditures. This insight has a high potential for further development of a model that can make the detection of infectious diseases.

**AUTHORS/INSTITUTIONS:** A. Radulovic, Institute of Public Health of Montenegro, Podgorica, Crna Gora, MONTENEGRO|

**CONTROL ID:** 3386757

**TITLE:** Validity and reliability of the patient safety culture questionnaire in hospital Naval Almirante Nef, Viña del Mar, Chile with confirmatory factor analysis and structural equations models

**ABSTRACT BODY:**

**Abstract Body:** The questionnaire measures the Patient Safety Culture (PSC) in hospitals. This instrument was constructed by the Agency for Healthcare Research and Quality (AHRQ). The University of Murcia, through the CUSEP project, translated its first version into Spanish.

The questionnaire permits to obtain benefits such as the reduction in the recurrence of patient safety incidents through notification and learning organization, reduction of the physical and psychological damage of patients based on preventing errors, optimization in resource management due to the effective risk assessment and constant changes in health care practices, among others. Additionally, systematic measurement allows to monitor continuously the strengths and weaknesses of the institution and its dependencies.

The spanish translation of the PSC questionnaire in Chile has not been validated yet. The strategy aims to applicate the spanish version for analyzing the performance of the questionnaire, in order to, afterwards, studying structure-related issues within the questions for adjusting them to the particular context of the country. This way, the CSP questionnaire would valid and reliable to apply.

The construct validity refers to the factorial structure of the measuring instrument. The degree of fit between the sample and the proposed hypothetical factor model is studied. For validating the CSP questionnaire, the considered methodology is the Confirmatory Factor Analysis (CFA) with structural equations models (SEM). Two types of rotation are carried out, the first supposes that the dimensions are orthogonal. On the other hand, the second supposes that the factors do not maintain independency structure.

Regarding the selection of the estimation method of factors it is carried out with principal factor axes, since it is intended to explain the greater variability of the data.

**AUTHORS/INSTITUTIONS:** K. Cuadros Carlesi, C.G. Leal Kaymaliz, Escuela de salud, Universidad Viña del Mar, Viña del Mar, Valparaíso, CHILE]

**CONTROL ID:** 3386762

**TITLE:** Modelling Gestational Difference of women in the Navrongo Municipality of the Upper East Region of Ghana.

**ABSTRACT BODY:**

**Abstract Body:** The achievement of Sustainable Development Goal three (3) which sought to reduce maternal mortality ratio to less than 70% per 100,000 live births by 2030 is threatened by the resurgence of unsteady Expected Date of Confinement (EDC) of pregnant women in Africa. EDC was used as the dependent variable coded with three categories: Below EDC (less than or equal to 37 weeks of gestation), within EDC (38 to 40 weeks of gestation) and above EDC (greater than or equal to 41 weeks of gestation).

The study used quadratic classification function analysis to investigate the influence of some characteristics of women on their gestational differences in the Navrongo municipality of the Upper East Region of Ghana. Retrospective data on some variables of delivered mothers and the neonates were extracted from the Biostatistics Unit of the War Memorial hospital in Navrongo. The extracted data spanned from January 2014 to the first month of 2017. The study excluded still-birth or macerated babies from its analysis. The results showed that, parity, age and baby's weight are the major discriminating variables in classifying women into classes of gestation in the Municipality. The study however revealed that parity was the most influential discriminating variable in classifying women into below EDC, Within EDC and above EDC of the period of gestation. The study recommended that women should treat every pregnancy differently and follow strictly the medical protocols during pregnancy. Research should also focus on possible biochemical explanations for this relationship, including the cellular mechanism behind parity and gestational difference.

**AUTHORS/INSTITUTIONS:** A. Alhassan, Statistics, University for Development Studies, Navrongo, GHANA|

**CONTROL ID:** 3386784

**TITLE:** Modified Simon's Minimax and Optimal Two-Stage Designs for Single-Arm Phase II Cancer Clinical Trials

**ABSTRACT BODY:**

**Abstract Body:** Simon's two-stage design and the admissible two-stage design have been commonly used in practice for single-arm phase II clinical trials when the primary endpoint is binary. The ethical benefit of the two-stage design over the single-stage design is attained by the early termination of the trial when the treatment seems to be inactive. While Simon's optimal design is the two-stage design that minimizes the expected number of subjects under the null hypothesis, the probability of falsely declaring futility after the first stage frequently seems undesirably high. In Simon's minimax design, however, it is often the case that a high proportion of the total planned subjects are evaluated in the first stage, and thus the ethical benefit may not be achieved. In this paper, we propose modified minimax and optimal two-stage designs which guarantee not only type I and II error rates but also reasonable sample size proportions in the first stage, while maintaining the probability of falsely declaring futility under a pre-selected level. The characteristics of the modified two-stage design will be compared with those of Simon's and the admissible two-stage design. The modified minimax design yields a design that requires modest increase in 29% of cases, while the modified optimal design saves 1 to 13 subjects in 81% of cases for  $\beta = 0.2$ . The modified design approach provides investigators with an alternative when the sample sizes of Simon's designs are severely unbalanced or the Type II error is unacceptably high after the first stage.

**AUTHORS/INSTITUTIONS:** J. Kim, M.J. Schell, Biostatistics and Bioinformatics, Moffitt Cancer Center & Research Institute, Tampa, Florida, UNITED STATES|J. Kim, Oncologic Sciences, University of South Florida, Tampa, Florida, UNITED STATES|

**CONTROL ID:** 3386854

**TITLE:** Multiple change-point estimation of panel data via EM algorithm

**ABSTRACT BODY:**

**Abstract Body:** This research addresses the problem of estimating multiple common change-points in multi-path panel data with the unknown number of change-points. The EM algorithm is used for maximization of the likelihood for mixture distributions. With the unknown number of change-points in parameters, the variable dimension of parameters causes computational difficulty in searching and resulting appropriate partitions. We suggest the tail-cutting algorithm which is recursive and repetitive as an alternative for binary search algorithm. We also extend our method to the multiple common change-points estimation to multi-path time series panel data. Simulations to evaluate the performance of the estimators are provided. Our method is applied to PM10 data in Republic of Korea

**AUTHORS/INSTITUTIONS:** J. Kim, Statistics, Duksung Women's University, Seoul, KOREA (THE REPUBLIC OF)|J. Kim, Statitics, Duksung Women's University, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3386868

**TITLE:** An improved keyboard design considering all toxicity information of candidate doses for phase I oncology trials

**ABSTRACT BODY:**

**Abstract Body:** The primary objective of phase I oncology trials is to determine the maximum tolerated dose (MTD), which is typically defined as the dose with a probability of dose-limiting toxicity that is closest to the target toxicity rate. The existing methods for phase I dose-finding designs are generally classified into algorithm-based designs, model-based designs, model-assisted designs. The algorithm-based designs are simple to implement but perform poorly for finding the MTD. The model-based designs are complex to implement but perform well. The model-assisted designs are simple to implement and perform well; however, the accumulated information for determining the dose escalation and de-escalation is not sufficiently used than the model-based designs.

A keyboard design, one of model-assisted designs, assumes a prior distribution of the probability of dose-limiting toxicity as an independent uniform distribution at each dose level. The prior distribution is individually updated by using the number of patients with a toxic response and the number of treated patients at each dose level to determine the dose escalation and de-escalation.

We improved the keyboard design so that all the accumulated information could be used to increase the precision of the MTD estimate. In this improved keyboard design, a dose–toxicity relationship is assumed for each dose level at the design stage of the trial and one plausible scenario is selected from the assumed dose–toxicity relationships by using the number of patients with a toxic response and the number of treated patients. The parameters of the prior distributions for each dose level are determined under the selected relationship and the prior distributions are updated in the same manner as the original keyboard design. The proportions of correctly selected MTD were higher when using the improved keyboard design than when using the original design.

**AUTHORS/INSTITUTIONS:** S. Ueyama, J. Tsuchida, S. Ando, T. Sozu, Tokyo University of Science, Katsushika, JAPAN|A. Hirakawa, The University of Tokyo, Bunkyo, JAPAN|

**CONTROL ID:** 3386869

**TITLE:** Variable selection for a propensity score model considering weak confounders

**ABSTRACT BODY:**

**Abstract Body:** In observational studies, propensity score (PS) methods are used to estimate an average treatment effect (ATE). An unbiased estimator of ATE is obtained when the PS model includes all confounders. Previous studies have illustrated that confounders have to be included in the PS model and that variables related only to the outcome (outcome predictors) should be also included in the PS model to increase the precision of the ATE estimator. In recent years, researchers often observe many variables and face difficulties in identifying the confounders and outcome predictors.

The outcome-adaptive lasso can be used to identify confounders and outcome predictors, because in this method, the penalty weight depends on the coefficients of the outcome regression model. However, confounders that have a weaker impact on the outcome than outcome predictors (weak confounders) may not be selected in the PS model. To tackle this issue, we propose a novel variable selection method directly considering the importance of the variables included in the PS model. The proposed method uses a penalty weight reflecting the coefficients of the outcome and treatment regression models. The penalty of confounders is smaller than that of the outcome predictors when confounders and outcome predictors have the same impact on the outcome. We compared the performance ((1) the bias, (2) the mean squared error of ATE estimators, and (3) the proportion of selected variables in the PS model) of the proposed method and the outcome-adaptive lasso, where the inverse probability weighted method was used to estimate the ATE. The bias and mean squared error of the proposed estimator was found to be smaller than that of the outcome-adaptive lasso.

**AUTHORS/INSTITUTIONS:** R. Fukushima, S. Ando, J. Tsuchida, T. Sozu, Tokyo University of Science, Katsushika, JAPAN|

**CONTROL ID:** 3386907

**TITLE:** Prediction Model Using Microbiome Data Successfully Distinguishes between Pancreatic Ductal Adenocarcinoma (PDAC) and Non-Cancerous Sample

**ABSTRACT BODY:**

**Abstract Body:** Pancreatic Ductal Adenocarcinoma (PDAC), the most common type of pancreatic cancer, is one of the deadliest cancer that shows poor prognosis. Most of PDAC patients are diagnosed with their disease in advanced stage, because PDAC is usually asymptomatic in early stage. Therefore, it is urgent to find a novel method for early detection of PDAC. While a few prediction biomarkers are available for early diagnosis such as CA19-9, its performance was not good. In this research we proposed using new types of biomarkers identified from extracellular vesicles (EVs) metagenome data to build a prediction model for early detection of PDAC. We used 87 PDAC (cases) samples from Seoul National University Hospital and 151 non-cancerous (control) samples from Baek Hospital and Boramae Hospital. However, it was hard to measure the pure effect of microbiota, because of significant confounding effects of sex and age between PDAC and non-cancerous samples. Thus, we selected subsamples using Propensity Score Matching (PSM) to reduce the cofounding effects of covariates. After matching, the confounding effects were greatly reduced. Using matched 50 cases and 67 control samples, several statistical methods were applied to find microbiota with differential abundance in phylum (L2) and genus (L6) levels between PDAC and normal samples. From these microbiota, we found best subset using an exhaustive search. Our prediction model showed that the test AUC was higher than 0.8 for both phylum (0.879) and genus levels (0.813), respectively. In summary, we developed a new prediction model for early detection of PDAC. The proposed model is based on microbiota and has high predictive power.

**AUTHORS/INSTITUTIONS:** K. Han, N. Kang, Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|J. Kim, J. Jang, H. Kim, Department of Surgery and Cancer Research Institute, Seoul National University College of Medicine, Seoul, KOREA (THE REPUBLIC OF)|T. Park, Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3386915

**TITLE:** Tailored software solutions for incorporating UAV for agronomy assessments

**ABSTRACT BODY:**

**Abstract Body:** Despite automated field phenotyping fast becoming an every day reality in agricultural research and production, many modern field agronomy assessments still rely almost exclusively on hand gathered sampling protocols for results. Modern solutions to age old problems of cost and logistic restrictions for field data collection provide a wide choice of options for meeting various cost and practical constraints. Drones are effective tools for collecting visual or spectral information, which can feed machine learning algorithms and computer simulations. Thanks to the support of high capacity external hard drives, SD cards, and long lived batteries, these methods can proceed in the field to open a new practical era of the promotion and implementation of advanced survey protocols for routine field sampling. Leveraging the flexibility of packages such as Shiny (Chang et al, 2018), a wider range of practitioners now have access to tailored software solutions, which encapsulate data capturing and complex statistical simulations, to promote cost effective and statistically robust field survey practices, without the requirement for expensive new hardware, or extensive training. In this talk, we discuss how it becomes possible to gather highly correlated auxiliary variables and use them to implement ranked set and judgement post-stratification sampling protocols, framed in the example of field seedling emergence assessment. We will also present the efficiency and cost-function analyses of those protocols in comparison to a simple random sample, showcasing a collaborative effort between biologists, statisticians and data scientists.

**AUTHORS/INSTITUTIONS:** P.J. Kasprzak, O. Ozturk, R.A. Edson, O. Kravchuk, Agriculture Food and Wine, University of Adelaide, Adelaide, South Australia, AUSTRALIA|O. Ozturk, Statistics, Ohio University, Columbus, Ohio, UNITED STATES|

**CONTROL ID:** 3386918

**TITLE:** Pathway analysis for Cross-over Design

**ABSTRACT BODY:**

**Abstract Body:** Cross-over designs have been widely used in clinical trials to investigate the efficacy of new treatments. In cross-over designs, each subject is treated subsequently with different treatments. Many methods such as linear mixed models have been used to analyze the repeated measurements from cross-over designs. When we consider repeated measured response variables, estimation of random components for linear mixed models is not always easy. Thus, we applied the generalized estimating equation (GEE) method to cross-over designs and compared their results with those from linear mixed models. To apply the GEE method to the data from the cross-over designs, we need to switch the role of variables such a way that the independent variable in linear mixed models is considered as a response variable in generalized estimating equation model and vice versa.

Based on this GEE framework we developed a novel pathway analysis for proteomics data generated from the cross-over design. Our main goal is to investigate associations between pathways and phenotypes in cross-over designs. Through simulation studies, we checked the type I errors and compared power to evaluate the performance of the proposed pathway method. Our proposed method was applied to the proteomics data from the cross-over design to investigate the effect of ingestion of seaweed extract on anti-oxidative and inflammation.

**AUTHORS/INSTITUTIONS:** M. Kamruzzaman, Y. Kim, T. Park, Department of Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|I. Huh, College of Nursing, Research Institute of Nursing Science, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|Y. Lim, O. Kwon, Department of Nutritional Science and Food Management, Ewha Womans University, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3386919

**TITLE:** Some remarks on bivariate distributions with Poisson structure

**ABSTRACT BODY:**

**Abstract Body:** Seshadri and Patil (1964) argued that one could not have a non-trivial distribution with  $X_1$  having a Poisson distribution and, for each  $x_2$ ,  $X_1 | X_2 = x_2$  also Poisson distributed. Specifically, there do not exist bivariate mass function such that both marginals and both conditionals being Poisson. In other respects, Arnold, Castillo and Sarabia (1999) consider the case in which, for each  $x_1$ ,  $X_2 | X_1 = x_1$  having a Poisson distribution, while for each  $x_2$ ,  $X_1 | X_2 = x_2$  also Poisson distribution. Such distributions are called Poisson conditionals distributions. They only have Poisson marginals in the case of independence. However, if we are satisfied with having one marginal  $(X_1)$  and the “other” family of conditionals  $(X_2 | X_1 = x_1)$  being of the Poisson form, then the Pseudo-Poisson distributions fill the bill precisely.

In the present note, we discuss distributional features of such models, characterization, explore inferential aspects and include an examples of an application of the Pseudo-Poisson model to the real data.

**AUTHORS/INSTITUTIONS:** M.B. Govinda Raju, School of Mathematics and Statistics, University of Hyderabad, Hyderabad, Telangana, INDIA|

**CONTROL ID:** 3386938

**TITLE:** The use of the discrete wavelet transformation in the detection of sleep stages using wrist accelerometry.

**ABSTRACT BODY:**

**Abstract Body:** Inadequate sleep is associated with substantial health and economic burden, and as such it is important that the amount and quality of sleep is measured in a simple and accurate manner for use in the general population. Accelerometry is a low-cost and noninvasive method that has been used to discriminate sleep from wake, however, its utility to detect sleep stages is unclear.

Using simultaneously measured accelerometry and polysomnography (the gold-standard in sleep detection) data from 242 healthy young adults (22 years old) in the Western Australian Pregnancy Cohort (Raine) Study we developed and compared methods which utilised raw, 30Hz, triaxial accelerometry data to classify stages of sleep.

A range of statistical and discrete wavelet transformation summary variables were developed to describe the raw accelerometry data. These variables were considered in differing combinations as observations in first- and second-order hidden Markov models (HMMs) with time-homogeneous and time-varying transition probability matrices. In addition, generalised linear mixed models (GLMMs) with multinomial responses and logit link functions were considered. Model predictions were compared to polysomnography and outcome accuracies and F-scores were calculated.

Consistently, HMMs yielded greater sleep stage detection than GLMMs, with time-varying HMMs yielding greater discrimination than time-homogeneous HMMs, but there was little difference between first- and second-order HMMs. HMMs produced the greatest F-scores when only statistical summary variables were considered. The addition of discrete wavelet transformation summary variables decreased F-scores, yet increased the minimum stage detection rates. Detection of sleep stages ranged 61-95%, whilst wake detection ranged 61-78%. These results suggest that wrist-worn accelerometry data may be able to detect sleep stages but that further investigation is required to optimise classification accuracy.

**AUTHORS/INSTITUTIONS:** M.L. Trevenen, B.A. Turlach, Mathematics and Statistics, University of Western Australia, Crawley, Western Australia, AUSTRALIA|P. Eastwood, K. Murray, University of Western Australia, Crawley, Western Australia, AUSTRALIA|L. Straker, Curtin University, Perth, Western Australia, AUSTRALIA|

**CONTROL ID:** 3387154

**TITLE:** Mass appraisal of land values using random forest with spatial restriction

**ABSTRACT BODY:**

**Abstract Body:** The advancement of computational software and machine learning practice has facilitated enhanced uptake of mass appraisal methodologies for price modelling and prediction of land value. Since the characteristics of properties are geographically distributed, spatial autocorrelation computing could improve models to explain property prices. Different types of Random Forest models (RF), the classical one and quantile RF (QRF), were recognized as machine learning technique for real estate mass appraisal. However, a major drawback of this method is that they ignore influences of neighboring observed data when predicting the price properties. In order to overcome the disadvantage, random forest plus kriging of residuals (RFKO) method can be used. Initially, a RF of land values using predictive ancillary variables is carried out in order to model the trend component. In the second step, ordinary kriging is applied to the residuals of RF and a spatial prediction of the residuals is created. The final prediction is an additive combination of both model steps. The aim of this study was to compare performances of RF and quantile QRF both with and without spatial restriction in the prediction of rural and urban land values. We use two datasets of 3718 and 264 market data, released between 2017 and 2018. The first contains data of rural land value for the whole Province of Córdoba, Argentina, and the second one involves data coming from a village (Villa María) in the Province of Córdoba. A 10-fold cross validation was used to estimate prediction errors for each model. The root mean square prediction error was expressed as percentage of the mean yield (RMSE). Additionally, we fit an empirical a theoretical semivariogram to characterize the Relative Structured Variability (RSV, ratio of nugget and sill variance) of the residual from the compared methods. The results showed that only in the urban land the methods that incorporate spatial information performed better, RMSE of 30% vs. 34% for RF and 33% vs. 34% for QRF with and without kriging of the residuals, respectively.

**AUTHORS/INSTITUTIONS:** M. Córdoba, M. Balzarini, National Scientific and Technical Research Council, Córdoba, Córdoba, ARGENTINA|F. Monzani, J. Carranza, M. Piumetto, Territorial Studies Center, National University of Córdoba, Córdoba, Córdoba, ARGENTINA|

**CONTROL ID:** 3387159

**TITLE:** Estimating the variance of the interaction term for enhancing replicability

**ABSTRACT BODY:**

**Abstract Body:** Lack of replicability has been of major concern in the past two decades in various fields of science (e.g. animal behavior, pre-clinical research and experimental psychology). In the field of animal behavioral studies, orchestrated multi-lab studies were held. Despite that, results were not replicated, proving that lack of replicability is the specific fallout of lab uncontrollable factors on each of the genotypes.

By modeling such a multi-lab study by a mixed 2-way ANOVA, where the lab effect is random and the studied effect is fixed, we can address the effects of the uncontrollable factors as the random interaction between the lab and the studied factor. By definition, an effect is replicable if it is significant in face of the increased variability. We, therefore, estimate interaction variance from multi-lab data, for the purpose of inference in a future related single-lab study. This goal puts the interaction variance estimation in the highlight. In this work, we devise estimation schemes to overcome the bias and increased MSE present in common estimation methods. The methods are extended to unbalanced sample size ANOVA, including missing combinations (cells, i.e. sample size 0). An experiment was designed and is conducted at 3 labs, to assess whether the approach fulfills its promises.

**AUTHORS/INSTITUTIONS:** I. Gosez, Department of Human Molecular Genetics and Biochemistry, Tel-Aviv University, Tel Aviv, Choose Any State/Province, ISRAEL|I. Jaljuli, Y. Benjamini, Statistics and Operations Research, Tel-Aviv University, Tel Aviv, ISRAEL|N. Kafkafi, I. Golani, Department of Zoology, Tel-Aviv University, Tel Aviv, ISRAEL|M. Bouge, E. Chesler, V. Philip, The Jackson Laboratory, Bar Harbor, Maine, UNITED STATES|

**CONTROL ID:** 3387165

**TITLE:** Network Analysis with Multi-Omics Data Using Graphical LASSO

**ABSTRACT BODY:**

**Abstract Body:** Precision matrix of variables provide information about their conditional dependencies and can be used to infer noble biological pathways or gene-protein interactions. Recently, as the cost of data generation has decreased, multiple biological omics datasets can be used in analyses to consider the specific structures of multi-omics data such as different sparsity of dependence within and between omics. In this study, we suggest a new method which can detect the interplay among multi-omics by using different penalization parameters based on graphical LASSO. The parameters can be determined by cross-validation to minimize the penalized likelihood function. The proposed method was evaluated with simulation data and showed that it successfully identified the disease susceptible multi-omics markers. The proposed method was applied to Chronic Obstructive Pulmonary Disease, which illustrates its practical value.

**AUTHORS/INSTITUTIONS:** J. Park, Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Seocho-gu, KOREA (THE REPUBLIC OF)|S. Won, Department of Public Health Sciences, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|S. Won, Institute of Health and Environment, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3387173

**TITLE:** Deep Learning Prognosis Model for Hepato Cellular Calcinoma

**ABSTRACT BODY:**

**Abstract Body:** Deep Learning Analytics uses predictive models that provide actionable information for HCC patient's better prognosis. It is a multidisciplinary approach based on HCC data processing, AI technology-learning enhancement, HCC neural network, and visualization. Three key components need further clarification to help them effectively apply deep learning in HCC prognosis to explain the methods for conducting deep learning, the benefits of using deep learning and the challenges of using learning analytics in HCC. Discover significant clinical factor and SNP markers to detect prognosis of HCC. Compare the efficiency with other prognosis model using support vector machine, liner discriminant, random forest, logistic regression by ROC curves. Using ICD-9 codes for HCC, 965 patients with HCC and all available data variables required to develop and test models were identified from a clinical and SNP records database. Data on 645 patients was utilized for development of the model and on 320 patients utilized to perform comparative analysis of the models. Clinical data such as presenting signs & symptoms, socio demographic data, presence of metastasis, laboratory data and corresponding diagnosis and outcomes were collected. Clinical data and SNP collected for each patient was utilized by to retrospectively ascertain optimal management for each patient. Clinical presentations and corresponding treatment was utilized as training examples..

**AUTHORS/INSTITUTIONS:** T. LEE, Data Science & Statistics, Korea National Open University, Seoul, KOREA (THE REPUBLIC OF)

**CONTROL ID:** 3387179

**TITLE:** An alternative approach to the estimation of biological age - avoiding the pitfalls of misinterpretation

**ABSTRACT BODY:**

**Abstract Body:** Given a set of age-related biomarkers, a biological age of an individual is conventionally defined as the average age of all individuals with a similar biomarker profile (the expected value of the age, conditional on the observed values of the biomarkers). Thus, one conventional way of predicting the biological age, is to fit a linear regression model for age, using the biomarkers as covariates. Differences of the individual's actual age from the predicted biological age would mean that the person is "biologically older" or "younger" than his/her real age, that is naturally interpreted as having shorter or longer residual life expectancy, respectively, compared to an average person of the same age. However, using a simple simulation study as well as the data of the Estonian Biobank, we will show that in many realistic cases such interpretation could be incorrect and highly misleading. If there are time-dependent risk factors that affect the biomarkers, younger or older "biological age" does not necessarily correspond to lower or higher hazard levels. To avoid such misinterpretation, we propose an alternative definition of the biological age, using the concepts of survival analysis and deriving the biological age predictions from a model for overall survival. Using the example of the Estonian Biobank, we compare both parametric (assuming the Gompertz distribution for survival time) and semiparametric estimation approaches with the conventional method.

**AUTHORS/INSTITUTIONS:** K. Fischer, A. Kolde, Institute of Mathematics and Statistics, University of Tartu, Tartu, ESTONIA|K. Fischer, A. Kolde, N. Taba, Institute of Genomics, University of Tartu, Tartu, ESTONIA|N. Taba, Institute of Molecular and Cell Biology, University of Tartu, Tartu, ESTONIA|E. Macdonald-Dunlop, P. Joshi, Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UNITED KINGDOM|

**CONTROL ID:** 3387185

**TITLE:** Utility of Multidimensional Scaling(MDS) and Causal Modeling in the Interpretation of fMRI Data

**ABSTRACT BODY:**

**Abstract Body:** Introduction: functional Magnetic Resonance Imaging(fMRI) is the most advanced technique to investigate the brain mechanism by measuring BOLD intensity. The data generated in fMRI studies are generally noisy, high dimensional, correlated, and complex fMRI data. There is great need for statistical methods to analyze fMRI data by splitting the variability into true and error components to establish group difference.

Objective: To integrate the fMRI with clinical and biological data of schizophrenia patients through MDS & causal modeling procedure

Methodology: Resting-state fMRI data of 40 schizophrenia patients and 40 healthy controls matched for age, sex, education, and handedness were recruited at National Institute of Mental Health and Neurosciences was used for the present study. Psychopathology of the patients has been assessed using SAPS &SANS. 116 Regions of Interest (ROI) measured over 150 time-points were extracted using MarsBar toolbox of SPM package. 51 regions that showed statistically significant difference between cases and controls at 5% level of significance were selected for further analysis.

MDS, a multivariate dimension reduction procedure is used to represent 51 ROIs in the lower dimensional space, such that the position of these regions in the reduced space is based on the distance matrix. Clustering of regions in the reduced space will combine the regions that show strong statistical dependency. Partition around Medoids algorithm is used in the reduced space and the influence of the obtained clusters on the positive and negative symptoms of schizophrenia was assessed using Path analysis.

A new approach for computing cluster activation score (CAS) has been proposed. Since BOLD intensities are likely to vary among different regions irrespective of disease conditions, a standardization is carried out to arrive at CAS. It is also intended to attach the discriminating ability of the corresponding regions while forming the CAS. Hence CAS was calculated by converting BOLD intensities in to quantiles of regions (Q) and t- statistics(T) as weights.

Results: MDS resulted 6 dimensions and 16 clusters. The SEM shows cluster which consists of Left superior temporal gyrus and left and right thalamus as a predictor of negative symptoms viz. Inattention( $\beta$ (SE)= 0.6(0.13); $p=0.01$ ); Apathy(0.43(0.17);0.03) and Anhedonia(0.6(0.26);0.01) of the schizophrenia.

**AUTHORS/INSTITUTIONS:** P.P. V, K. Thennarasu, S.D. D K, Biostatistics, National Institute of Mental Health and Neurosciences, Bengaluru, INDIA|V. G, Psychiatry, National Institute of Mental Health and Neurosciences, Bangalore, Karnataka, INDIA|

**CONTROL ID:** 3387192

**TITLE:** An extension of ridge regression model on fuzzy data

**ABSTRACT BODY:**

**Abstract Body:** Ridge regression model is a widely used model with many successful applications, especially in managing correlated covariates in a multiple regression model. Multicollinearity represents a serious threat in fuzzy regression models as well. We address this issue by combining ridge regression with the fuzzy regression model. Our proposed algorithm uses the  $\alpha$ -level estimation method to evaluate the parameters of the ridge regression model with fuzzy data. Simulation experiments and an empirical study with crisp independent variables and a fuzzy dependent variable are presented. Results show that the proposed model can reduce the effect of multicollinearity across a wide spectrum of spreads for the fuzzy response, from moderate to very high levels of correlation between input variables.

**AUTHORS/INSTITUTIONS:** H. Jung, Faculty of Liberal Education , Seoul national university, Seoul, KOREA (THE REPUBLIC OF)|H. Kim, Department of Statistics, North Carolina State University, Raleigh, North Carolina, UNITED STATES|

**CONTROL ID:** 3387197

**TITLE:** Multi-study factor analysis to derive shared and study-specific dietary patterns in the United States (INHANCE consortium)

**ABSTRACT BODY:**

**Abstract Body:** Dietary patterns (DPs) are important tools for describing multi-dimensional dietary information through a small number of variables. With the increasing evidence on the role of empirically derived DPs – based on multivariate statistical methods - in the risk of several diseases, it becomes crucial to examine the extent to which similar DPs are consistently seen across centers from the same study or across different studies, potentially representing different populations or countries. A few papers considered this aspect so far; they mostly separately identified DPs across studies and assessed potential similarities with elementary statistics. However, when individual-level data from the single studies are available, an integrated statistical model could outperform previous analyses and reveal unknown similarities between DPs derived across different studies.

Following this direction, we have recently proposed in nutritional epidemiology the use of multi-study factor analysis (MSFA), a novel approach designed to simultaneously learn DPs common to all the available studies, as well as study-specific DPs present in some studies only. We applied MSFA to 7 case-control studies, from Europe and the United States, participating in the International Head and Neck Cancer Epidemiology (INHANCE) consortium. We identified 3 DPs shared among the 7 studies and 1 study-specific DP for each of the American studies (giving 4 study-specific DPs in total). The 4 American study-specific DPs showed a similar factor-loading matrix, which opposed calcium and niacin.

To further investigate similarities between the study-specific DPs, we propose to apply MSFA on the subset of the previous data including the 4 American studies only. We identified 3 DPs shared across all the US studies and 2 study-specific patterns, one for the Memorial Sloan Kettering Cancer Center and the other for the North Carolina (2002-2006) study.

We have made a step forward in understanding the extent to which DPs are reproducible across different studies from the same country.

**AUTHORS/INSTITUTIONS:** R. De Vito, Biostatistics, Brown University, Providence, Rhode Island, UNITED STATES|V. Carla, DEPARTMENT OF CLINICAL SCIENCES AND COMMUNITY HEALTH, UNIVERSITA' DEGLI STUDI DI MILANO, Milano, Milano, ITALY|

**CONTROL ID:** 3387209

**TITLE:** Modelling and Estimation of Multivariate Causal Associations in Health

**ABSTRACT BODY:**

**Abstract Body:** In biomedical studies, often researchers collect multiple outcomes from an individual and the outcomes could be correlated. The interest would be to estimate the effect of an intervention or exposure on these outcomes. Apart from simultaneous modelling of the association between the intervention and the outcomes, one would be interested to ascertain if the joint association is causal to demonstrate to policy-makers the effectiveness of the intervention on disease occurrence. Most studies with multiple outcomes have simply analysed them separately using standard causal inference statistics. Limited Statistical methodology developed for ascertaining causal association in repeated measured outcomes is adopted here to develop methods for assessing the causal association between an exposure and a multivariate outcome. The proposed method is evaluated by a simulation study and its application ascertained by an analysis of real-life data on the effect of exclusive breastfeeding on child trivariate nutritional outcome: stunting, wasting and underweight in Malawi.

**AUTHORS/INSTITUTIONS:** H.S. Twabi, Mathematical Sciences, University of Malawi, Zomba, MALAWI|S.O. Manda, Biostatistics Research Unit, South African Medical Research Council, Pretoria, Gauteng, SOUTH AFRICA|S.O. Manda, Department of Statistics, University of Pretoria, Pretoria, SOUTH AFRICA|D. Small, Department of Statistics, University of Pennsylvania, Pennsylvania, Pennsylvania, UNITED STATES|

**CONTROL ID:** 3387226

**TITLE:** Hierarchical structural component models for pathway analysis using RNA sequencing data

**ABSTRACT BODY:**

**Abstract Body:** In the recent years, technical improvements and decreasing costs of next generation sequencing technology (NGS) made RNA sequencing an alternative to the microarray. Many numbers of methods and software were proposed for the identification of differentially expressed genes (DEGs). However, analyzing high-throughput gene expression data at the pathway level can be effective rather than just identifying a DEGs. Identifying active pathways that differ between the two conditions can have more explanatory power than simple list of genes. Several analyses for identifying cancer-associated pathways based on gene expression data are mostly based on single pathway analyses, and thus do not consider correlations between pathways.

In this study, we propose a hierarchical structural component model for pathway analysis using RNA sequencing data for binary phenotype. Our method accounts for the hierarchical structure of genes and pathways in the single model while considering the correlations among pathways simultaneously. Our method has the ability to perform conditional inference for identifying a novel pathway given pathways. In application to a real biological data analysis, we demonstrated that our method could successfully identify pathways associated with the diagnosis of cancer.

**AUTHORS/INSTITUTIONS:** L. Mok, Interdisciplinary Program in Bioinformatics , Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|T. Park, Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|S. Lee, Center for Precision Medicine, Seoul National University Hospital, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3387228

**TITLE:** Dynamic Networks for non-Gaussian Time-series Data

**ABSTRACT BODY:**

**Abstract Body:** In this talk, I will introduce sparse dynamic chain graph models for network inference in high dimensional non-Gaussian time series data. The proposed methodology parametrized by (i) a precision matrix that encodes the intra-time conditional independence interactions among variables at each time-point (partial correlation networks), and (ii) an autoregressive coefficient matrix that contains dynamic conditional independences networks among time series components across consecutive time steps (Granger causality). In this work, we assume a stable dynamic chain graph model meaning that the structure of interactions within each time point remains stable for previous and current time step, and interactions between consecutive time points are stable too.

The proposed model is a Gaussian copula vector autoregressive model, which is used to model sparse interactions in a high-dimensional setting. Estimation is achieved via a penalized EM algorithm. In this paper, we use an efficient coordinate descent algorithm to optimize the penalized log-likelihood with the smoothly clipped absolute deviation penalty. We demonstrate our approach on simulated and genomic datasets. The method is implemented in an R-package entitled tsnetwork.

**AUTHORS/INSTITUTIONS:** P. Behrouzi, Wageningen University and Research, Utrecht, NETHERLANDS|F. Abegaz, University of Groningen, Groningen, NETHERLANDS|E. Wit, University of Liège, Liège, SWAZILAND|

**CONTROL ID:** 3387242

**TITLE:** Propensity Weighting in the Estimation of Direct Treatment Effects.

**ABSTRACT BODY:**

**Abstract Body:** We investigate propensity weighting in assessing direct effects in a model where treatment may be mediated by another risk factor. In particular, we compare two approaches to estimating direct effects, counterfactual approaches as discussed by Vanderweele (2015) and principal stratification as suggested by Rubin (2002, 2004). We demonstrate the ideas via simulation studies and apply the method to a study of Cancer of Unknown Primary (CUP). The exposure variable is Cancer of Unknown Primary (CUP) and the mediator is treatment. A CUP is confirmed after a set of recommended tests are performed but a primary cancer is not found. A CUP is unconfirmed if the tests are not done. The outcome is survival beyond a fixed time point. A direct effect of a CUP diagnosis would suggest that CUP is directly affecting survival, irrespective of treatment. An indirect effect would be survival by modified by the treatment

**AUTHORS/INSTITUTIONS:** C. Drake, Department of Statistics, University of California, Davis, California, UNITED STATES|J. Smith-Gagen, Epidemiology, University of Nevada, Reno, Nevada, UNITED STATES|

**CONTROL ID:** 3387249

**TITLE:** Effect of population stratification on SNP-by-environment interaction.

**ABSTRACT BODY:**

**Abstract Body:** Proportions of false-positive rates in genome-wide association analysis are affected by population stratification, and if it is not correctly adjusted, the statistical analysis can produce the large false-negative finding. Therefore various approaches have been proposed to adjust such problems in genome-wide association studies. However, in spite of its importance, a few studies have been conducted in genome-wide single nucleotide polymorphism (SNP)-by-environment interaction studies. In this report, we illustrate in which scenarios can lead to the false-positive rates in association mapping and approach to maintaining the overall type-1 error rate.

**AUTHORS/INSTITUTIONS:** J. An, S. Won, Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|S. Won, Interdisciplinary Program for Bioinformatics, College of Natural Science, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|S. Won, Institute of Health and Environment, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|C. Lange, S. Won, Department of Biostatistics, Harvard T. H. Chan School of Public Health, BOSTON, Massachusetts, UNITED STATES|C. Lange, Channing Division of Network Medicine, Brigham and Women's Hospital, BOSTON, Massachusetts, UNITED STATES|

**CONTROL ID:** 3387250

**TITLE:** Clustering of trajectories based on multistate model

**ABSTRACT BODY:**

**Abstract Body:** Clustering of sequential or temporal data is challenging especially when dealing with discrete data. Our motivating problem derive from the need to find patterns of drug use trajectories over time.

It is essential to have standard measures of change, define appropriate similarities among trajectory observations, obtain appropriate data representation and use methods that are suitable for this kind of data or using information extracted from them in order to apply classical methods.

We analysed data from a sample of 70000 drug users to identify transition patterns in drug use trajectories across 5 year.

Data were collected every three months for a total of 20 measurements for each subject, demographic and some clinical covariates are available.

Preliminary analysis was done using optimal matching and the three-step procedure proposed by Leffondree et al. (2004) to identify clusters of individual longitudinal trajectories.

We propose to addressed the problem of unsupervised classification of sequences using a multistate approach, in order to obtain measures to quantify change in drug use behaviours; these approach allow to consider also the effects of covariates. These set of measures could be used as input for classical clustering methods but could also help to provide effective visualizations for applied use.

The obtained results will be compared with the ones obtained by the other proposed methods.

**AUTHORS/INSTITUTIONS:** R. Miglio, Statistical Sciences, University of Bologna, Bologna, ITALY]

**CONTROL ID:** 3387265

**TITLE:** Latent Growth Trajectories of Depression in the Perinatal Period

**ABSTRACT BODY:**

**Abstract Body:** Background: Pregnancy and Postpartum period are strenuous on both the mental and the physical health of a woman. Depression over the perinatal period and its growth trajectory is influenced by the unobserved heterogeneity amongst expecting mothers. A growth mixture modeling approach can be used to identify distinct latent classes (LC) and trajectories of depression.

Methodology: Repeated measurements on the Edinburgh Postnatal Depression Scale (EPDS) at three trimester of pregnancy, at five weeks, 6 months, 12 months and 24 months postpartum for the women in the BCHADS Study (a cohort of 909 pregnant women in Bangalore) were of interest. A Latent Class Growth Analysis (LCGA) was performed for EPDS to explore the number of LCs with distinct growth trajectories for depression in perinatal period. Unconditional Growth Mixture Model (GMM) was used to finalize the number of LCs. A conditional GMM with predictors: baseline Violence, Stress and Life Events was fitted to estimate their effect on LCs.

Result: In the antenatal period, A four LC solution using LCGA was found to be statistically relevant albeit the unconditional GMM identified a three LC to be optimum. The three visibly distinct trajectories (growth parameters) were of the Normative Group (NG), Sharply Decreasing Depression (SDD), and Sharply Increasing Depression (SID) with 88.14%, 6.71%, & 5.14% of individuals in each subgroup respectively. Violence, Stress, and Life Events were significant risk factor for baseline depression levels in SDD group when compared with NG (OR; 95% CI: 1.65; 1.169-2.33, 1.42; 1.155-1.75, and 1.6; 1.052-2.454). However, Stress and Life Events were risks for the SID group when compared with NG (OR; 95% CI: 1.27; 1.058-1.524 and 1.6; 1.106-2.338).

Conclusion: GMM can be used effectively to identify hidden heterogeneity in a seemingly homogenous population. Depression levels have a differential growth process over the course of pregnancy and postpartum period. They are affected by baseline covariates such as Violence and Stress.

Keywords: Latent Class Growth Analysis, Growth Mixture Modeling, EPDS, Antenatal, Perinatal.

**AUTHORS/INSTITUTIONS:** A. Bajaj, Biostatistics, National Institute of Mental Health and Neurosciences, Bengaluru, Karnataka, INDIA|K. Thennarasu, Biostatistics, NIMHANS, Bengaluru, Bengaluru, INDIA|G. Desai, P.S. Chandra, Psychiatry, NIMHANS, Bangalore, Karnataka, INDIA|A. Pickels, Biostatistics, Kings College London, London, UNITED KINGDOM|H. Sharp, University of Liverpool, Liverpool, UNITED KINGDOM|

**CONTROL ID:** 3387270

**TITLE:** Determination of Diet influence on Type 2 Diabetes using Hierarchical Structure Component Model(HisCoM)

**ABSTRACT BODY:**

**Abstract Body:** Abstract—Type 2 diabetes(T2D) makes about 90% of cases of diabetes and according to World Health Organisation, the number of people diagnosed with T2D is on the rise annually even among young people. The development of T2D is caused by a combination of lifestyle and genetic factors. While some of these factors are under personal control, such as diet and obesity, other factors are not, such as age, sex, and genetics. This study used the Korea Association Resource (KARE) study dataset with 4,506 samples (T2D cases, controls) to determine the association between diet and T2D. In our preliminary analysis, original Recommended Food Score(RFS) in the dataset; which is defined as a total sum of 46 food items did not show any significant effects of diet on T2D. Original RFS treats all food items equally by using the same weight. In this study, we then looked for new methods for optimal weights for the 46 food items which in turn can maximize the correlation between T2D and diet. Two methods, Principal

Component Analysis(PCA) and model based analysis were then considered to calculate new weights for the food items. These weights were then used to compute new weighted RFSs. Between these two new weighted RFSs, one that maximizes the relationship between T2D was chosen as the optimal weighted RFS and was used with other covariates to determine the influence of diet on the T2D.

Key words: type 2 diabetes, recommended food score, principal component analysis, predictor

**AUTHORS/INSTITUTIONS:** C. APIO, T. Park, DEPARTMENT OF BIOINFORMATICS AND BIostatISTICS, SEOUL NATIONAL UNIVERSITY, Seoul, Gwanak-gu , 1 Gwanak-ro, KOREA (THE REPUBLIC OF)

**CONTROL ID:** 3387274

**TITLE:** Comparison of Batch Effect Elimination Methods in Liquid Chromatography/Time-of-Flight Mass Spectrometry

**ABSTRACT BODY:**

**Abstract Body:** The batch effect is frequently observed in untargeted liquid chromatography/time-of-flight mass spectrometry (LC-MS) metabolomics in an analysis of large samples due to instrumental variation. To increase statistical power and reproducibility of a statistical model based on metabolic data, batch effect elimination is needed. In this work, we compare batch effect elimination methods: Batch normalizer, Support vector regression, QC-normalization, Support vector regression, and QC-normalization, and Systematic Error Removal using Random Forest (SERRF). We calculate the relative standard deviation (RSD) in the quality control pooled sample (QC) to assess the performance of those methods in Korean-based metabolic data.

Key words: Metabolomics, Batch effect, LC-MS, Batch normalizer, Support vector regression, QC-normalization, Systematic Error Removal using Random Forest

**AUTHORS/INSTITUTIONS:** T. Jung, Seoul National University, Gwangmyeongsi, KOREA (THE REPUBLIC OF)|T. Park, Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3387278

**TITLE:** Using Dimension Reduction Strategy to Combine Clinical and Proteomic Data: An Applied Study of Alzheimer's Disease

**ABSTRACT BODY:**

**Abstract Body:** Alzheimer's disease is one of the most common neurocognitive disorder and affects quality of life in elderly. Therefore, accurate diagnosis of the disease is crucial in practice. As known, high dimensional molecular data have become popular and widely used to classify patients and predict disease status in biostatistics and biomedicine for more than a decade. In addition, researchers examine the methods to combine proteomic information with the clinical data such as patient's age, gender, family history, comorbidity, etc. to improve diagnostic accuracy. There are several different strategies introduced in the literature to use both clinical and proteomic data for classification. However, it is important to evaluate and validate high-dimensional molecular data's contribution to the performance of the prediction model [1,2].

In this study, we aim to investigate the effect of dimension reduction strategy on classification performance in elderly patients with Alzheimer's disease. We use both clinical and proteomics data of 194 patients. In first step principal component analysis (PCA) is used for reducing the number of proteins in proteomics data. Then, we used reduced version of the proteomics and combine them with clinical variables to classify patients with random forest (RF) and support vector machine (SVM) with hold-out validation approach. In addition, we examine the predictive performance of models using both clinical and proteomics data with models using clinical and proteomics data, separately to validate contribution of combining clinical data with proteomics in classification performance.

It was seen that even though classification performance of PCA+SVM is slightly higher than the PCA+RF approach, test performance of PCA+RF were higher in terms of both accuracy, sensitivity, specificity in all the models.

[1] Boulesteix, A. L., & Sauerbrei, W. (2011). Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in bioinformatics*, 12(3), 215-229. [2] Little, R.J.A., and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, second edition. Wiley, New York.

[2] Tibshirani, R. J., & Efron, B. (2002). Pre-validation and inference in microarrays. *Statistical applications in genetics and molecular biology*, 1(1).

**AUTHORS/INSTITUTIONS:** S. Karahan, H.Y. Zengin, Biostatistics, Hacettepe University, Ankara, TURKEY|Y. Ayhan, Psychiatry, Hacettepe University, Ankara, TURKEY|E. Saka Topçuoğlu, Neurology, Hacettepe University, Ankara, TURKEY|A.T. Baykal, Biochemistry, Acibadem University, Istanbul, TURKEY|E. Özkan, Social Sciences and Humanities, Koç University, Istanbul, TURKEY|B. Sahin, Acibadem Labmed R&D Center, Istanbul, TURKEY|

**CONTROL ID:** 3387288

**TITLE:** Willingness to pay for Pro-Vitamin A Crop Varieties:the case of Yellow maize in Ghana

**ABSTRACT BODY:**

**Abstract Body:** Food security inevitably go with improved nutrition to guarantee a healthy and active life. Poor meal even though diversified is often associated with deficiency in micronutrient and it is a strong indicator for child stunting and maternal nutritional status. Yellow and orange maize contain high levels of  $\beta$ -carotene, making it an important crop for combating Vitamin A deficiency (VAD) which is prevalent among children and pregnant women in Ghana. CSIR-CRI in collaboration with the HarvestPlus has developed several maize varieties that contain appreciable levels of pro-vitamin A in Ghana. Intensive dissemination of these vitamin a maize varieties is currently on-going. A study was conducted to ascertain the acceptability of these new maize varieties for cultivation and potential willingness to pay for these varieties in Ghana. Using a sample of 227 farmers from two major maize producing regions and employing bidding elicitation technique, willingness to pay was assessed. Results revealed that farmers had low WTP of about GHC3.00 (USD 0.50) per kg of orange maize seed. Price, knowledge, age, experience and education affected WTP significantly. It is therefore suggested that interested stakeholders come together to set acceptable price for increased production.

Key words; bidding elicitation, Ghana, micronutrient, Nutrition, Price

**AUTHORS/INSTITUTIONS:** P.P. Acheampong, L. Brobbey, Socioeconomics, CSIR-Crops Research Institute, Kumasi, GHANA|M. Ewool, A. oppong, P. Ribiero, Cereals, CSIR-Crops Research Institute, Kuamsi, GHANA|M. Mochiah, Plant Health, CSIR-Crops Research Institute, Kumasi, GHANA|E.A. Obeng, Biometrics, CSIR-Crops Research Institute, Kumasi, GHANA|

**CONTROL ID:** 3387347

**TITLE:** Prediction and classification of kidney transplant patient evolution based on longitudinal functional metabolomics data

**ABSTRACT BODY:**

**Abstract Body:** Omics datasets are longitudinally measured to detect changes in profiles predicting outcomes. Our work is motivated by a biomarker study of repeatedly measured metabolomics dataset in 18 kidney transplant patients. The aim of this study is to predict and classify 18 kidney transplant patients based on their NMR spectra taken from blood samples. Since the NMR spectra of each patient at a fixed time point are functional data and the data are collected up to nine time points during and after surgery, we propose longitudinal functional data analysis methods to summarize our data and a novel clustering method for curves to achieve our aim. We extracted four feature curves and further eight feature values from each patient based on longitudinal functional principal component analysis (FPCA). These feature curves and values are used for classification. Specifically, we do two dimensions FPCA, one is on spectra dimension and the other one is on time dimension. From the dense spectra direction, we extracted four feature curves (functions of time) which are the FPC scores corresponding to the first four FPCs. These score curves, explaining a large proportion of the variability of the data, are extracted features which can be used for classification. Furthermore, in order to extract simpler features, we conducted FPCA on these extracted score curves along time and in total eight scores values for each patient (two from each score curve) are extracted. These score values (each patient has eight values) can also be used for classification. We propose a classification based on extracted feature curves and feature values by developing a novel nonparametric supervised and unsupervised functional classification methods. Lastly, we compare our method with classification method on the data directly via simulations.

**AUTHORS/INSTITUTIONS:** M. Xie, X. Long, H. Liu, J. Houwing-Duistermaat, Department of Statistics, University of Leeds, Leeds, UNITED KINGDOM|E. Paci, School of Molecular and Cellular Biology, University of Leeds, Leeds, UNITED KINGDOM|

**CONTROL ID:** 3387357

**TITLE:** Finite Mixtures of Multiple Scaled Unrestricted Skew-Normal Generalized-Hyperbolic Distribution: Bayesian Approach

**ABSTRACT BODY:**

**Abstract Body:** The multiple scaled unrestricted Skew-Normal Generalized-Hyperbolic (MS-SUNGH) is an attractive and flexible family of probability distributions, which can provide different degrees of heavy-tailedness and asymmetric properties for each dimension of the variable space. This family contains, as special or limiting cases, many symmetric and asymmetric distributions, such as Scale Mixtures of Normal (SMN), Scale Mixtures of Skew-Normal (SMSN), symmetric Generalized-Hyperbolic (GH) and Skew-Normal Generalized-Hyperbolic (SNGH). This paper presents a Bayesian approach to estimation of the MS-SUNGH family. In particular, we develop an MCMC approach using Gibbs sampling which has some advantages in this context as most of the Gibbs sampling updates are available in closed form. The approach is illustrated, and the performance examined, by applying the MS-SUNGH to mixture model problems using simulated and real data in medicine and biology which present challenging clustering problems.

**AUTHORS/INSTITUTIONS:** D. Wraith, School of Public Health & Social Work, QUT, Deagon, Queensland, AUSTRALIA|M. Maleki, Statistics, Shiraz University, Shiraz, IRAN (THE ISLAMIC REPUBLIC OF)

**CONTROL ID:** 3387409

**TITLE:** Constrained functional additive models for estimating interactions between a treatment and functional covariates

**ABSTRACT BODY:**

**Abstract Body:** A novel functional additive model is proposed which is uniquely modified and constrained to model nonlinear interactions between a treatment indicator and a potentially large number of functional/scalar covariates. We generalize functional additive regression models by incorporating treatment-specific components into additive effect components. A structural constraint is imposed on the treatment-specific components, to give a class of orthogonal main and interaction effect additive models. If primary interest is in interactions, we can avoid estimating main effects, obviating the need to specify their form and thereby avoiding the issue of model misspecification. The methods are illustrated with data from a clinical trial with imaging data as predictors.

**AUTHORS/INSTITUTIONS:** H.G. Park, E. Petkova, T. Tarpey, Department of Population Health, New York University, New York, New York, UNITED STATES|T. Ogden, Biostatistics, Columbia University, New York, New York, UNITED STATES|

**CONTROL ID:** 3387430

**TITLE:** Title: A Cost-Utility Analysis of HIV-related Malnutrition In Women Using A Bayesian Probabilistic Approach With A Markov Chain Decision Tree Model

**ABSTRACT BODY:**

**Abstract Body:** Objectives

This study investigates the emergence of the triple threat – HIV, Malnutrition and Drug Resistance in women aged 15 – 49 years of age. The primary purpose of this study is to model the prognosis of a HIV – positive female ART patient to assist in the decision-making analysis of the effectiveness of a nutritional intervention program.

Design and Methods

This study investigates the emergence of the triple threat – HIV, Malnutrition and Drug Resistance in women aged 15 – 49 years of age. Using a Bayesian probabilistic approach with a Markov Chain decision tree model in a resource-limited health care provider setting, sequential algorithms in different transition health states were derived using utility values extrapolated from a systematic review. This cost-utility analysis allowed for the calculation of the Quality Adjusted Life Years (QALYs) gained by the patient from the introduction of this intervention and compared with the WHO-Choice threshold of gross domestic product per capita.

Results

The expected cost of a nutritional intervention and no nutritional intervention in conjunction with an ART regime in female HIV/AIDS patients aged 15 – 49 years were \$US 1129.31 and \$US 181.64, respectively. Additionally, the QALYs gained were 13.56 and 2.87 for a nutritional intervention and no nutritional intervention, respectively. The incremental cost per QALY gained for the nutritional intervention versus no nutritional intervention was \$ 10.68 per QALY and is notably well below the WHO's annual threshold for developing countries in the Caribbean.

Conclusion

The introduction of a nutritional program to supplement HIV-positive female ART patients of reproductive ages is exceedingly cost-effective and gives good value for money. Such an intervention should be an integral component in the treatment regime for HIV-positive patients.

**AUTHORS/INSTITUTIONS:** V. Sankar, Mathematics and Statistics, The University of the West Indies, St. Augustine, Valsayn, TRINIDAD AND TOBAGO|

**CONTROL ID:** 3387432

**TITLE:** Modelling Human Mobility Patterns and Social Contacts for Predicting Infectious Disease Spread

**ABSTRACT BODY:**

**Abstract Body:** Human movement and social contacts play a key role in the spread of infectious diseases such as influenza and measles among others. For an infectious contact to occur, there must be social contact which is facilitated by individuals that leave their home locations for different reasons such as work or school. Various models have been proposed which purport to capture the average travel behaviour of individuals using underlying population density measures and are routinely used to predict movement patterns in epidemic forecast models. However, these models rarely distinguish between different types of individuals, nor do they incorporate individual and population-level heterogeneities in travel behaviour. Here, we proposed a model of movement which explicitly models an individual's multiple trips to destinations. Random effects to capture the heterogeneities in individual travel patterns were introduced in the modelling framework. We applied the model to data collected over the first year of a longitudinal cohort study on mobility and social contacts set in and around Guangzhou Province, China. Results show that majority of the contacts were made in the vicinity of the home locations, within a radius of 6km from home with key differences by gender, employment status and location (whether rural or urban). The model with key individual characteristics consistently performed better than those modelling the average flow of individuals. The model indicates the inclusion of individual-level characteristics better captures mobility patterns. This knowledge can play a key role in the design of non-pharmaceutical interventions such as movement restriction for effective control against the spread of infectious diseases.

**AUTHORS/INSTITUTIONS:** J. Chirombo, Statistical Support Unit, Malawi-Liverpool-Wellcome Trust, Blantyre, MALAWI|P.J. DIGGLE, J.M. Read, CHICAS, Lancaster University, Lancaster, Lancashire, UNITED KINGDOM|D.J. Terlouw, Malaria Epidemiology Group, Malawi-Liverpool-Wellcome Trust, Blantyre, MALAWI|

**CONTROL ID:** 3387445

**TITLE:** A family of nonparametric and semi-parametric estimators to diagnostic accuracy when diagnostic tests are subject to detection limits---Application to diagnosing Alzheimer's disease

**ABSTRACT BODY:**

**Abstract Body:**

In medical diagnosis, subjects are usually assumed to be one of two basic types ----- healthy and diseased (the gold standard). Often times, candidate diagnostic markers are much cheaper and less invasive than the gold standard, but must be compared to the gold standard in their sensitivity and specificity to accurately diagnose the diseases. When candidate diagnostic markers are fully measured, Receiver Operating Characteristic (ROC) curves have been the standard approaches for measuring the diagnostic accuracy. However, measurements of diagnostic markers may not be available above or below some limits due to various practical and technical limitations. For example, in the diagnosis of Alzheimer disease (AD) using cerebrospinal fluid (CSF) biomarkers, the Roche Elecsys® immunoassays have a measuring range for multiple CSF molecular concentration. Many cognitive tests used in diagnosing AD are also subject to floor and ceiling effects. We propose a novel statistical methodology for estimating the diagnostic accuracy when a diagnostic marker is subject to detection limits by dividing the fully observed measurements into two parts by a threshold. We propose a family of estimators to the area under ROC curve (AUC) with only minimum parametric assumptions by combining a conditional nonparametric estimator and another conditional semi-parametric estimator derived from a Cox's proportional hazards model. We derive the variance to the proposed estimators, and further assess the performance of the proposed estimators as a function of possible thresholds through an extensive simulation study, and recommend the optimum thresholds. Finally, we apply the proposed methodology to assess the ability of a range of CSF biomarkers and cognitive tests in diagnosing AD using real world data from Washington University Knight Alzheimer Disease Research Center.

**AUTHORS/INSTITUTIONS:** C. Xiong, Biostatistics, Washington University in St. Louis, St. Louis, Missouri, UNITED STATES|J. Luo, Public Health Sciences, Washington University in St. Louis, St. Louis, Missouri, UNITED STATES|J. Morris, Neurology, Washington University in St. Louis , St. Louis, Missouri, UNITED STATES|

**CONTROL ID:** 3387448

**TITLE:** A Machine Learning-Based Approach to Cancer Classification using RNA-Seq Data

**ABSTRACT BODY:**

**Abstract Body:** In the recent past, machine-learning approaches have gained a lot of attention in the biomedical field mainly for biological classification. A number of researchers are currently applying these methods to the RNA-Seq data. RNA-Seq technology typically generates a huge amount of data making the search for useful genes in any given study a daunting task. Differentially expressed gene list usually yields a large list of genes even after adjusting for multiple testing. This can make subsequent studies quite cumbersome and extensive. In this study, we evaluate the applicability of machine learning approaches to classify colorectal cancer patients into either being in the early or late cancer stages based on the differentially expressed genes. We use publicly available colorectal cancer RNA-Seq dataset extracted from the TCGA database. The final dataset consists of 273 samples (154 in early-stage, 119 in late-stage). The sample is divided into training and test set. Four commonly used machine-learning methods are applied to the datasets and their performances evaluated using three model evaluation metrics; accuracy, Kappa and Area Under the Curve (AUC). The results show that the Support Vector Machine (SVM) can be a better choice of a classification method for cancer patients using RNA-Seq data.

Keywords: RNA-Seq, Machine Learning, AUC, Differentially Expressed Genes, Colorectal Cancer

**AUTHORS/INSTITUTIONS:** L. Chaba, Institute of Mathematical Sciences, Strathmore University, Nairobi, Nairobi, KENYA|B. Omolo, Mathematics and Computer Science, University of South Carolina - Upstate, Spartanburg, South Carolina, UNITED STATES|

**CONTROL ID:** 3387454

**TITLE:** Modelling threshold levels on Pest & Diseases in sugarcane management

**ABSTRACT BODY:**

**Abstract Body:** Decisions on whether to eradicate sugarcane grown in a farmer's field are based on a pre-set threshold level of pest and disease infestation. We conduct a review of the existing approach that includes sampling procedures and hypothesis testing method. The use of standard t-test in estimating the level of pest and disease infestation ignores the variability due to different inspectors and field management thus provides underestimates of the disease infestation rates. We propose an alternative approach to address the problem. Using a two stage systematic sampling method on sugarcane out growers' fields, different inspectors inspect for pest and diseases infection on different varieties at different maturity age. A total sample of 587 observations (percent disease infection) were used in setting threshold levels that are compared to standard threshold levels in use. We introduce a mixed model approach in drawing inference as to whether to eradicate the infested sugarcane field or not. Upon adjustment for the variety, age and interaction effects, we compare the estimated mean of disease infestation to pre-set threshold levels and construct a 95 % confidence intervals of the mean infestation. The standard errors used incorporate the field, inspector within field and correlation effect. The sampling procedure in use provides sufficient samples that are unbiased and representative of the true pest and disease infestation. The variety, age and variety by age interaction effect were insignificant. We propose use of confidence interval decisions that are based on the mixed model approach because it adjust for the variety, age and interaction effects while incorporating variability in the fields, inspectors within fields effect. Adopting this approach will lead to fewer rejects and invariability, a cost reduction in farmer production.

**AUTHORS/INSTITUTIONS:** P. Njuho, Department of Statistics, University of South Africa, Johannesburg, Gauteng, SOUTH AFRICA|N.O. Ama, Statistics, University of Botswana, Gaborone, BOTSWANA|N. Sewpersad, Biometrics Unit, South Africa Sugar Research Institute, Durban, KwaZuluNatal, SOUTH AFRICA|

**CONTROL ID:** 3387457

**TITLE:** The MELODIC family for simultaneous binary logistic regression of multiple outcome variables in a reduced space

**ABSTRACT BODY:**

**Abstract Body:** Logistic regression is a commonly used method for binary classification. Oftentimes, researchers have more than a single binary response variable and simultaneous analysis is beneficial because it provides insight into the dependencies among response variables as well as between the predictor variables and the responses. In this paper we propose the MELODIC family for simultaneous binary logistic regression modeling. In this family the regression models are defined in an Euclidean space of reduced dimension, based on a distance rule. The model may be interpreted in terms of logistic regression coefficients or in terms of a biplot. We discuss a fast MM algorithm for parameter estimation. Two applications are shown in detail: one relating personality characteristics to drug consumption profiles and one relating personality characteristics to depressive and anxiety disorders.

**AUTHORS/INSTITUTIONS:** M.D. Rooij, Methodology and Statistics, Leiden University, Leiden, NETHERLANDS|

**CONTROL ID:** 3387458

**TITLE:** A joint model of multiple longitudinal and multiple time-to-event traits for genetic studies

**ABSTRACT BODY:**

**Abstract Body:** Longitudinal studies are arguably most suitable to determine direct &/or indirect association of genetic variants (SNPs) with time-to-event traits while accounting for dependences with intermediate risk factors. Although various models are available for analysis of multiple longitudinal traits, multiple time-to-event outcomes, and joint single longitudinal & time-to-event; less attention has been paid to joint models (JM) for multiple longitudinal & multiple time-to-event traits. Here we formulate a novel JM that consists of: (i) a linear mixed model for the longitudinal traits that describes the trajectory of each trait as a function of time, SNP effects and includes subject random effects to account for dependences within/between traits; (ii) a frailty Cox PH model for the time-to-event outcomes that depends on SNP & trajectory effects from (i) and includes a subject frailty term to account for the dependence between traits. Models (i) & (ii) are fitted sequentially along with a bootstrap estimate of the joint covariance matrix. Using generalized Wald tests, we classify a SNP association with a time-to-event trait as indirect if SNP effect is non-null in (i) & null in (ii); direct if SNP effect is null in (i) & non-null in (ii); and direct & indirect if SNP effects are non-null in (i) & (ii). We also propose a multi-trait test for global SNP association with all (or a subset) of the traits. Motivated by the genetic architecture of Type-1 diabetes complications observed in the Diabetes Control and Complications Trial (DCCT), we demonstrate feasibility by an application in DCCT & assess model performance by a realistic simulation study under a scenario based on DCCT. This scenario involves 5 causal SNPs with direct effects on two time-to-complications (retinopathy, nephropathy) &/or indirect effects via two observed (HbA1c, blood pressure) and one simulated longitudinal risk factors. Compared to standard approaches (separate analysis of each trait, time-to-event analysis adjusted on longitudinal trait values), our approach provides better performance (lower bias, Type-I error control & relative power) and correctly retrieve direct/indirect SNP associations. JM of multiple longitudinal & multiple time-to-event outcomes can provide insight into etiology of multifactorial traits while accounting for potential adverse effects of measurement errors & reverse causality.

**AUTHORS/INSTITUTIONS:** M. Brossard, O. Espin-Garcia, S.B. Bull, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, CANADA|A.D. Paterson, Hospital for Sick Children Research Institute, Toronto, Ontario, CANADA|A.D. Paterson, O. Espin-Garcia, S.B. Bull, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, CANADA|R. Craiu, Statistical Sciences, University of Toronto, Toronto, Ontario, CANADA|

**CONTROL ID:** 3387465

**TITLE:** The use of multivariate latent class mixed effect models to predict pathological fetal growth restriction

**ABSTRACT BODY:**

**Abstract Body:** Background

Fetal growth restriction (FGR) is defined as an estimated fetal weight (EFW) of < 10<sup>th</sup> percentile and is associated with short- and long-term adverse perinatal outcomes. The International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) recommends Doppler ultrasound of the umbilical artery (UA) and middle cerebral artery (MCA) to separate fetuses with pathological FGR (FGR-P), who are at higher risk for adverse outcomes, from constitutionally small fetuses, considered to have non-pathological FGR (FGR-C).

**Aim**

To use latent class modeling of prenatal ultrasound data to identify pathological and non-pathological subject clusters, compare them to groups based on the ISUOG criteria, and assess differences in infant body composition measurements.

**Methods**

Longitudinal fetal ultrasound measures and infant outcomes from 79 participants in the Perelman Family Foundation FGR study were included. Fetuses were classified as FGR-C or FGR-P based on ISUOG criteria using the last prenatal ultrasound. Multivariate latent class mixed effects modeling with longitudinal EFW, UA and MCA was also used to classify fetuses. Differences in group classification based on the ISUOG criteria and modeling were examined. Infant birth outcomes and postnatal body composition measures were compared between the groups.

**Results**

ISUOG classification yielded groups of 38 FGR-P and 41 FGR-C fetuses. Our best fit model ultimately delineated 2 groups based on EFW and MCA. Of the 55 infants identified as pathological based on our model, 23 were classified as FGR-C and of the 24 infants identified as non-pathological by our model, 6 were classified as FGR-P. The model groups differed significantly in several infant characteristics at birth, similar to differences between the ISUOG groups. There were significant differences between model groups in 3 postnatal body composition outcomes: HC:weight ratio, mid-upper arm circumference and quadriceps skinfold.

**Conclusion**

Multivariate latent class mixed effects modeling yielded similar groupings of infants as standard clinical criteria and differentiated infants in terms of postnatal body composition.

**AUTHORS/INSTITUTIONS:** C. Palmer, C. Driver, L.D. Brown, J. Hobbins, L. Pyle, Pediatrics, University of Colorado School of Medicine, Aurora, Colorado, UNITED STATES|L. Pyle, Biostatistics and Informatics, Colorado School of Public Health, Aurora, Colorado, UNITED STATES|

**CONTROL ID:** 3387467

**TITLE:** Bivariate traits association analysis using generalized estimating equations in family data

Mariza de Andrade<sup>1</sup>, Mauricio A. Mazo Lopera<sup>2</sup>, Nubia E. Duarte<sup>3</sup>

<sup>1</sup> Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA ; <sup>2</sup> Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia; <sup>3</sup> Department de Matemáticas, Universidad Nacional de Colombia, Manizales, Colombia

**ABSTRACT BODY:**

**Abstract Body:** Genome wide association (GWAS) is becoming fundamental in the arduous task of deciphering the etiology of complex diseases. The majority of the statistical models used to address the genes-disease association consider a single response variable. However, it is common for certain diseases to have correlated phenotypes such as in cardiovascular diseases. Usually, GWAS typically sample unrelated individuals from a population and the shared familial risk factors are not investigated. In this paper, we propose to apply a bivariate model using family data that associates two phenotypes with a genetic region. Using generalized estimation equations (GEE), we model two phenotypes, either discrete, continuous or a mixture of them, as a function of genetic variables and other important covariates. We incorporate the kinship relationships into the working matrix extended to a bivariate analysis. The estimation method and the joint gene-set effect in both phenotypes are developed in this work. We also evaluate the proposed methodology with simulation study and an application to real data.

**AUTHORS/INSTITUTIONS:** M. de Andrade, Health Sciences Research, Mayo Clinic, Rochester, Minnesota, UNITED STATES|

**CONTROL ID:** 3387471

**TITLE:** A Weibull-Gompertz Makeham Distribution with Properties and Application to Cancer Data

**ABSTRACT BODY:**

**Abstract Body:** The article presents an extension of the Gompertz Makeham distribution using the Weibull-G family of continuous probability distributions proposed by Tahir et al. (2016a). This new extension generates a more flexible model called Weibull-Gompertz Makeham distribution. Some statistical properties of the distribution which include the moments, survival function, hazard function and distribution of order statistics were derived and discussed. The parameters were estimated by the method of maximum likelihood and the distribution was applied to a bladder cancer data. Weibull-Gompertz Makeham distribution performed best (AIC = -6.8677, CAIC = -6.3759, BIC = 7.3924) when compared with other existing distributions of the same family to model bladder cancer data.

**AUTHORS/INSTITUTIONS:** P.O. Koleoso, A.U. Chukwu, Statistics, University of Ibadan, Ibadan, Oyo State, Ibadan, Oyo State, NIGERIA|

**CONTROL ID:** 3387488

**TITLE:** MULTIVARIATE ANALYSIS BASED ON GENETIC DISSIMILARITY MEASURES WITH BIFACTORIAL DESIGN

**ABSTRACT BODY:**

**Abstract Body:** Genetic variability between molecular haplotypes of groups of individuals is measured multivariately via binary molecular markers. Multivariate variation can be studied through dissimilarity measures. A geometric partitioning of multivariate variation, through distance metrics that consider the nature of molecular data, based on analysis of variance design, allows us to obtain significant values (p-values) using non-parametric distribution based on ranks. The sums of squares between and within groups of individuals, based on molecular analysis of variance (AMOVA), can be inferred from the matrix of distances between pairs of haplotypes. The classical AMOVA allows the evaluation of statistical significance between two or more factors in a nested design, i.e. hierarchical factors. Consequently, it is not possible to evaluate the interaction between factors. The interaction occurs when all the levels of a factor have been evaluated in each of the levels of the other factor. The objective of this work was to present a method to evaluate the interaction between factors and the main effect with binary data. The statistical inference of this AMOVA under a non-hierarchical model is based on a non-parametric test based on ranks proposed by Kruskal Wallis. The proposed analysis of the main effect and interaction term in a non-hierarchical AMOVA includes: calculation of the distance matrix and its partition into blocks, subsequent calculation of residuals and non-parametric analysis of variance on the residuals. This method can select different distance metrics. We tested the presented method using a simulation experiment on a 2×2 factorial design. The results suggest that the proposed test for the Non-Hierarchical AMOVA has high power both for quantifying the interaction and for evaluating the main effects when the interaction is not statistically significant.

**AUTHORS/INSTITUTIONS:** C.I. Bruno, Biometric and Statistics, National University of Córdoba, Córdoba, Córdoba, ARGENTINA|M.E. Videla, CONICET - UNC - UNVM, Córdoba, Córdoba, ARGENTINA|M. Balzarini, UNIVERSIDAD NACIONAL DE CÓRDOBA, Cordoba, Cordoba, ARGENTINA|

**CONTROL ID:** 3387513

**TITLE:** Novel Clustering Methodology for Investigating Patient Heterogeneity in Disease Course and Underlying Immunological Biomarkers

**ABSTRACT BODY:**

**Abstract Body:** To improve the understanding of the human immune system in rheumatic conditions, it is necessary to identify up-stream immunological biomarkers or signatures that relate to disease impact (e.g. remission, progression) or response to therapy.. In particular, identifying subpopulations of patients with particular immunological profiles linked to clinical outcomes would help to not only improve our understanding of the biology but potentially help to identify drug targets and improve management and treatment of patients. Under this stratified or precision medicine perspective, we develop a probabilistic outcome-driven clustering approach based on Dirichlet mixture modelling of a latent disease phenotype, derived from latent class modelling of a longitudinal clinical outcome reflecting disease impact, and the immunological markers. The approach is an extension of the 'Profile Regression' approach of Molitor et al. (2010) and is applied to 263 rheumatoid arthritis patients where disease activity is measured repeatedly over time up to a maximum follow-up of 18 months and baseline data on 163 autoantibodies directed against human autoantigens in autoimmune disease were collected.

**AUTHORS/INSTITUTIONS:** B.D. Tom, MRC Biostatistics Unit, University of Cambridge, Cambridge, UNITED KINGDOM|

**CONTROL ID:** 3387515

**TITLE:** Sequential basket trial design based on multi-source exchangeability with predictive probability monitoring

**ABSTRACT BODY:**

**Abstract Body:** Precision medicine endeavors to conform therapeutic interventions to the individuals being treated and needs to account for the heterogeneity of treatment benefit among patients and patient subpopulations. Basket trials comprise a class of experimental designs that endeavor to test the effectiveness of a therapeutic strategy among patients defined by the presence of a particular biomarker target rather than a particular cancer type, where the evaluation of treatment effectiveness are conducted with respect to the “baskets” which collectively represent a partition of the targeted patient population. We present novel methodology for a sequential basket trial design with Bayesian interim analyses based on a novel hierarchical modeling strategy for sharing information among a collection of discrete, potentially non-exchangeable subtypes. Based upon simulation studies to elucidate the statistical properties, we illustrate that the proposed design leads to potential gains in trial efficiency relative to more traditional two-stage designs that do not facilitate subgroup detection. For example, both family-wise and marginal type I error rates can be reduced, decreasing the probability of a false positive, while potentially reducing the expected sample size. Additionally, the power can be increased in certain scenarios while maintaining control of the type I error rate. There is also an increased probability of terminating the study early due to futility or efficacy based on the proposed design and methods as compared to standard designs, which can save and better allocate resources. The proposed design illustrates the potential for improved efficiency of basket trial designs to examine the effectiveness of treatment in multiple subgroups with sequential monitoring.

**AUTHORS/INSTITUTIONS:** A. Kaizer, Biostatistics and Informatics, University of Colorado-Anschutz Medical Campus, Aurora, Colorado, UNITED STATES|N. Chen, MD Anderson Cancer Center, Houston, Texas, UNITED STATES|J. Koopmeiners, University of Minnesota-Twin Cities, Minneapolis, Minnesota, UNITED STATES|B. Hobbs, Cleveland Clinic, Cleveland, Ohio, UNITED STATES|

**CONTROL ID:** 3387547

**TITLE:** Risk factors identification with spatially correlated survival data with multiple units per spatial coordinate

**ABSTRACT BODY:**

**Abstract Body:**

Spatially correlated right censored data are encountered in many fields such as biomedical studies, migration of individuals, epidemiology, actuarial science to name a few. Suppose we have  $k$  locations that are pairwise spatially correlated and that  $n_i, i=1, \dots, k$  individuals are in location  $i$ . The units in each of the areas are monitored for the occurrence of an event. In this talk, we present a new class of models for failure time data that combine both modern survival analysis and geostatistical formulation. Regression-type model parameters for risk factors identification as well those of a Matern-type spatial correlation are obtained via composite likelihood. Possible association between units in a given spatial location is embedded in the models. Large and small sample properties will be presented. An application for models illustration will be discussed. The models can be applied to the situation where the areas of interest are modeled for the occurrence of an event

**AUTHORS/INSTITUTIONS:** A. Adekpedjou, Z. Sainul, Mathematics and Statistics, University of Missouri-Science & Technology, Rolla, Missouri, UNITED STATES|

**CONTROL ID:** 3387553

**TITLE:** Clustering gene expression and other health related data using Bayesian intrinsic dimension

**ABSTRACT BODY:**

**Abstract Body:** A Bayesian non parametric model is developed to identify the data intrinsic dimensions which can be seen as the number of relevant variables that are needed to completely describe the data-generating process, i.e. the dimension of all the nonredundant information contained in a dataset. The methodology is used in the Schmitz dataset, a collection of more than 25000 rna-sequenced genes measured for 564 biopsy samples affected by diffuse large B-cell lymphoma, or DLBCL. Three clusters are identified that show significantly different Kaplan-Meier survival functions. Similar results are obtained in the Golub microarray dataset, of leukemia patients. The same methodology used on fMRI and protein folding data unveils interesting data structures starting from a pure geometrical prospective.

**AUTHORS/INSTITUTIONS:** A. Mira, Computer Science, Università della Svizzera italiana, Lugano, SWITZERLAND|  
A. Mira, Università dell'Insubria, Como, ITALY|

**CONTROL ID:** 3387554

**TITLE:** Prognosis prediction with Multi-omics data  
for Pancreatic Ductal Adenocarcinoma

**ABSTRACT BODY:**

**Abstract Body:** Pancreatic Ductal Adenocarcinoma (PDAC), which accounts for more than 90% of patients with pancreatic cancer, has a very poor prognosis with a 5-year overall survival rate less than 8%. Therefore, identifying prognostic biomarkers and improving prognosis are central goals for PDAC research. Although there have been several approaches proposed for prognosis of PDAC, none have provided high accuracy enough for clinical application. Most of these approaches have used only one type of omics data. We produced Whole Exome Sequencing (WES) and RNA-seq data for 196 patients with PDAC at Seoul National University Hospital (SNUH) to discover biomarkers related with prognosis. Somatic mutations and Copy Number Variation (CNV) were called using the WES data for each sample. In this study, we performed an integrative analysis of multi-omics data to discover biomarkers using genome and transcriptome data and to build model for predicting prognosis. We considered model based approaches such as penalized Cox regression and machine learning approaches including Survival Analysis Learning with Multi-Omics Neural network (SALMON) model. We adopted a validation set approach to compare the performance of these approaches. We first constructed the prediction models with identified biomarkers using SNUH hospital data and then validate them using an independent PDAC data with 150 patients from The Cancer Genome Atlas (TCGA). We used evaluation measures such as C-index.

Key words: PDAC, survival analysis, multi-omics, somatic mutation, copy number variation, RNA-seq, neural network

**AUTHORS/INSTITUTIONS:** T. Goo, Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|S. Jeong, T. Park, Department of Statistics, Seoul National University, Seoul, KOREA (THE REPUBLIC OF)|D. Park, Department of Life Sciences, Ajou University, Suwon, KOREA (THE REPUBLIC OF)|J. Jang, Department of Surgery and Cancer Research Institute, Seoul National University College of Medicine, Seoul, KOREA (THE REPUBLIC OF)|

**CONTROL ID:** 3387563

**TITLE:** Area-under-the-curve metrics for describing patient-reported toxicity data in cancer clinical trials

**ABSTRACT BODY:**

**Abstract Body:** In oncology, the changing landscape towards chronically administered therapies has resulted in a call by stakeholders for enhanced descriptions of patient tolerability in clinical trials (Thanarajasingam G, et al. Lancet Haematol. 2018). To date, a variety of longitudinal graphics and summary metrics have been developed to describe clinician-reported adverse event (AE) data in the ToxT macro suite in SAS (Thanarajasingam G, et al. Lancet Oncol. 2016), though questions remain unanswered about optimality of individual graphics and summary metrics. The Patient-Reported Outcomes version of the Common Terminology Criteria for AEs is a library of questions developed by the US National Cancer Institute which can be used to develop custom patient surveys to improve the accuracy of symptomatic AE reporting in oncology clinical trials. The aim of this project was to develop area-under-the-curve (AUC) based longitudinal graphics and summary metrics to describe PRO-CTCAE data. The resulting graphics and metrics were subsequently applied to PRO-CTCAE data collected in a metastatic prostate cancer phase III trial. AUCs were defined at the patient- and group-level with and without taking into account baseline (i.e., pre-existing) symptom levels (i.e., "incremental" AUCs). Patient-level AUCs were defined with and without proration to account for varying lengths of time on treatment. Longitudinal graphics included a longitudinal plot of treatment arm means (based on mixed model estimation) with shading to represent the group-level AUC. Graphical depiction of patient-level AUCs relied on exploration of the distribution of individual values via histograms and box plots. A comparison of AUC summary metrics for PRO-CTCAE items as well as the range of longitudinal graphics when applied to clinical trial data will be presented. AUC metrics for the summarization of PRO-CTCAE data appear as interpretable for a variety of symptomatic AEs in practice and merit further investigation. Care must be taken to describe computation method as variability was observed across possible definitions.

**AUTHORS/INSTITUTIONS:** A.C. Dueck, P. Novotny, G.L. Mazza, Health Sciences Research, Mayo Clinic, Scottsdale, Arizona, UNITED STATES|L. Rogak, E. Basch, Memorial Sloan Kettering Cancer Center, New York, New York, UNITED STATES|E. Basch, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina, UNITED STATES|G. Thanarajasingam, Division of Hematology, Mayo Clinic, Rochester, Minnesota, UNITED STATES|

**CONTROL ID:** 3387581

**TITLE:** THE IMPLEMENTATION OF VIRTUAL LEARNING OBJECTS AS A DIDACTIC STRATEGY FOR THE TEACHING OF THE CONCEPT OF PROBABILITY

**ABSTRACT BODY:**

**Abstract Body:** The teachers of basic, middle and higher education have the challenge of addressing issues that are evident in everything that surrounds us, nowadays globalization forces them to analyze and conclude on all kinds of information that reflects the behavior of the economy of a country, of the world, of the climate change, crop behavior, or data in any area of the knowlag. That is why in this research a didactic sequence is being used in the framework of meaningful learning, strengthened by the empirical analytical approach, and the design of the qualitative research process Participative Action Research; to build a Virtual Learning Object-OVA-, in order to strengthen Random Thinking and Data Systems - PASD-, in particular for the concept of probability, given that these themes are directly supported by concepts and procedures of the theory of probabilities and inferential statistics, and indirectly in descriptive and combinatorial statistics; in order to strengthen the PASD in teachers students in the Master in Mathematics Didactics of the Faculty of Distance Studies of the Pedagogical and Technological University of Colombia-UPTC-.

The Ministry of National Education (2006) has implemented five thoughts in the Colombian themathematics curriculum, including the PASD, among the modules of the Master's curriculum is the Mathematical Thought Module, in which the PASD is addressed, and the teacher teachers comment that due to lack of knowledge of the subjects they do not risk addressing them; stating that they did not have both disciplinary and didactic knowledge (didactic knowledge of the content) of statistics and probability, being conscious that these issues are transversal in all areas of the know.

Considering this problem, the following question was raised that is guiding the research process:

How would the design and application of OVAs be achieved to strengthen and strengthen the PASD in the teacher teachers of the Master's in Didactic of Mathematics at UPTC?

To answer it, the effect of the OVAs on the appropriation of the concept of probability is being established, in the students teachers of the Master, who have a future can enhance their learning in students of learning basic and academic middle.

**AUTHORS/INSTITUTIONS:** V.M. BURBANO PANTOJA, ESCUELA DE MATEMATICAS Y ESTADISTICA, UNIVERSIDAD PEDAGOGICA Y TECNOLOGICA DE COLOMBIA, Tunja, COLOMBIA|M.A. VALDIVIESO MIRANDA , V.M. BURBANO PANTOJA, PAR EVALUADOR COLCIENCIAS, Bogotá, COLOMBIA|V.M. BURBANO PANTOJA, INVESTIGADOR ASOCIADO COLCIENCIAS, Tunja, COLOMBIA|M.A. VALDIVIESO MIRANDA, MATEMATICAS Y ESTADISTICA, UNIVERSIDAD PEDAGOGICA Y TECNOLOGICA DE COLOMBIA, Tunja, COLOMBIA|M.A. VALDIVIESO MIRANDA, INVESTIGADOR JUNIOR COLCIENCIAS, TUNJA, COLOMBIA|

**CONTROL ID:** 3387602

**TITLE:** New insights towards a robust prediction model approach for sugarcane traits based on NIR spectra

**ABSTRACT BODY:**

**Abstract Body:** Sugarcane researchers are investing in breeding programs targeting new varieties that may supply food and bioenergy industry needs. On breeding programs, genomic selection (GS) is considered as an option to increase selection accuracy. However, the implementation of GS in polyploid species, such as sugarcane, is still a challenge. Recently, breeders are trying to integrate high-throughput phenotyping (HTP) information into GS. Researchers are exploiting HTP platforms to provide phenotypic records that can be either treated as secondary traits or as predictor variables together with molecular markers in a single-trait analysis. Breeders must find phenotyping methods that are fast, exact, consistent, and easy to apply. In this sense, Near-Infrared Spectroscopy (NIR) comes into play. NIR has been proven to be useful for prediction purposes. This work aimed to investigate how well SNP markers can capture NIR spectra variability in a sugarcane dataset. Also, this information will be considered to pursue better and more robust phenotypic prediction models. NIR spectra and SNP-based information were obtained for 385 clones under selection for fiber (FB) and apparent sucrose (PC) content. NIR spectra were obtained from dry powder bagasse. The NIR data was pre-treated, following the recommendation from the chemometrics experts, aiming at the increase of signal to noise ratio. We tested different pre-treatments: mean centering and scaling, smoothing with Savitzky-Golay algorithm, multiplicative scatter correction, and first and second derivatives. Models were fitted using the Bayesian ridge regression approach. The average prediction accuracy was obtained from 20 cross-validations, with 80 and 20% of the samples for the training and testing population. Genomic heritability ( $h^2$ ) was highly variable across the spectra. We found  $h^2$  peaks above 60%, showing the existence of strong polygenic signals for some wavelengths. The identification of wavelengths more associated with polygenic signals is dependent on the pre-treatment. Higher prediction accuracy occurred when wavelengths associated with the highest  $h^2$  were included in the models. We may consider this approach for wavelength selection, aiming to create more straightforward, informative, and robust prediction models for sugarcane. (CAPES, CNPq)

**AUTHORS/INSTITUTIONS:** L.A. Peternelli, M. Goncalves, Estatística, Universidade Federal de Viçosa, Vicosa, Minas Gerais, BRAZIL|G. Morota, Animal and Poultry Science, Virginia Tech, Blacksburg, Virginia, UNITED STATES|

**CONTROL ID:** 3387629

**TITLE:** Combining phenotypes to reduce multiple testing burden in PheWAS using ICD codes

**ABSTRACT BODY:**

**Abstract Body:** Phenome-wide association studies (PheWAS) is a study design that scans important genetic variants over thousands of phenotypes. For example, once a significant genetic variant is identified by Genome-wide Association Study (GWAS), association between this variant and numerous disease phenotypes are tested. One popular way to explore PheWAS is via international classification of disease (ICD) codes in electronic health records (EHR). One of the main challenges in PheWAS studies is to correct for the multiple testing when phenotypes are correlated. To minimize correlation and reduce the burden of multiple testing, we explore different ways to combine correlated phenotypes into a summary score.

**AUTHORS/INSTITUTIONS:** K. Kim, Preventive Medicine, Northwestern University, Chicago, Illinois, UNITED STATES|

**CONTROL ID:** 3387650

**TITLE:** Using Termite Mound Point Patterns from Drone and Aerial Data Acquisition to Understand Ecosystem Disturbance.

**ABSTRACT BODY:**

**Abstract Body:** During the exploration of natural resources, the environment and sensitive ecosystems can be disturbed. In ecology, indicator species have been used as a proxy to understand how sensitive ecosystems are to these disturbances. Disturbances can either be natural or anthropogenic. This study, used termite mounds as indicator species to monitor natural and anthropogenic ecosystem disturbance in the Eastern Cape Karoo South Africa, ahead of potential Shale Gas development. Termite mound spatial data was collected using three techniques, mainly drone, ground and aerial data acquisition. The two remote surveying techniques were compared against ground-surveyed data to understand which one is more optimal. In order to understand the effects of natural and anthropogenic disturbance on mounds, data on geology layer type, human settlements, elevation, vegetation and soil were collected. To establish the optimal surveying technique, termite mound point patterns were analysed for consistency with the ground-surveyed data. Point patterns from termite mound distributions informed by natural and anthropogenic disturbance were also analysed. Clark and Evans Aggregation Index R index, as well as the G-function, were applied.

**AUTHORS/INSTITUTIONS:** L.S. Mngcele, Geosciences, Nelson mandela university, Port Elizabeth, Eastern Cape, SOUTH AFRICA|L.S. Mngcele, African Earth Observatory Network, Port Elizabeth, Eastern Cape, SOUTH AFRICA|

**CONTROL ID:** 3387982

**TITLE:** Multiplicative rates model for recurrent events in case-cohort studies

**ABSTRACT BODY:**

**Abstract Body:** In large prospective cohort studies, accumulation of covariate information and follow-up data make up the majority of the cost involved in the study. This might lead to the study being infeasible when there are some expensive variables and/or the event is rare. Prentice (*Biometrika* 73(1):1–11, 1986) proposed the case-cohort study for time to event data to tackle this problem. There has been extensive research on the analysis of univariate and clustered failure time data, where the clusters are formed among different individuals under case-cohort sampling scheme. However, recurrent event data are quite common in biomedical and public health research. In this paper, we propose case-cohort sampling schemes for recurrent events. We consider a multiplicative rates model for the recurrent events and propose a weighted estimating equations approach for parameter estimation. We show that the estimators are consistent and asymptotically normally distributed. The proposed estimator performed well in finite samples in our simulation studies. For illustration purposes, we examined the association between prior occurrence of measles on acute lower respiratory tract infections (ALRI) among young children in Brazil.

**AUTHORS/INSTITUTIONS:** J. Cai, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, UNITED STATES|L. Amorim, Federal University of Bahia, Salvador, BRAZIL|P. Maitra, GlaxoSmithKline Pharmaceutical company, Philadelphia, Pennsylvania, UNITED STATES|

**CONTROL ID:** 3387986

**TITLE:** Bayesian Inference for Extreme Outcome Dependent Sampling Design

**ABSTRACT BODY:**

**Abstract Body:** Outcome Dependent Sampling (ODS) designs, such as the case-control and case-cohort designs, have been widely used for their study efficiency. In this paper, we propose a new and cost-effective sampling design, the extreme outcome dependent sampling (EODS) design, for studies with a continuous outcome. Compared to existing ODS designs, the new EODS design adopts a different strategy to use the smallest and largest outcomes to identify supplemental samples. The EODS design provides an alternative way to make efficient use of limited resources, especially in multi-state studies, by targeting the more informative subjects for sampling. We develop a Bayesian Markov Chain Monte Carlo (MCMC) method for the EODS design. Our method can incorporate the information of all subjects, including those with unobserved exposure, and provide an unbiased and efficient estimator through posterior inference. Simulation results indicate that the newly proposed EODS scheme, coupled with the proposed MCMC estimator, is more efficient than the existing ODS designs and the simple random sampling design with the same sample size. The proposed method is illustrated with a data set from the Collaborative Perinatal Project.

**AUTHORS/INSTITUTIONS:** H. Zhou, T. Wang, G. Koch, Biostatistics, Univ. of North Carolina at Chapel Hill, Chapel Hill, North Carolina, UNITED STATES|X. Wang, Biostatistics & Bioinformatics, Duke University, Durham, North Carolina, UNITED STATES|

**CONTROL ID:** 3390194

**TITLE:** Modelling flowering time of wheat using accumulated heat with GAMs

**ABSTRACT BODY:**

**Abstract Body:** Modelling plant growth from time of sowing to flowering is an important decision aid for determining when to grow cereal grain cultivars in Australia. Growing different cereal cultivars that make optimal use of environment conditions while avoiding water-deficiency at the end of the growing season is essential for Western Australian agriculture. Knowing approximate flowering dates of wheat, barley, and oat cultivars allows growers to manage late seasonal stresses such as frost and water-stress. When the cultivars flower is dependent on time-of-sowing, location, season and the genetic properties of the cultivar. Research undertaken by Sharma et al (2011) used accumulated temperature from time of sowing as a predictor of flowering date. We extended this research with more recent datasets and utilised a penalised spline estimator to fit generalised additive models (GAMs) to take into account genetic by environment by farm-management interactions. This presentation will provide insights to modelling trial-datasets and extending these models to predict locations and time-of-sowing combinations for which data is not available. A large proportion of our research is focused on ensuring that the predictions are accurate. Our research also provides a workflow to ensure the best possible predictions are achieved.

**AUTHORS/INSTITUTIONS:** D. Diepeveen, K. Ng, K. Salam, K. Reeves, Department of Primary Industries and Regional Development, Government of Western Australia, Perth, Western Australia, AUSTRALIA|D. Diepeveen, Veterinary and Life Sciences, Murdoch University, Perth, Western Australia, AUSTRALIA|K. Reeves, Curtin University, Bentley, Western Australia, AUSTRALIA|

**CONTROL ID:** 3390229

**TITLE:**

Assesing the robustness of the estimators of the Cumulative incidence function for Semi and Non-parametric methods in the analysis of competing risk data

**ABSTRACT BODY:**

**Abstract Body:**

The aim of this study is to investigate the properties of the estimators of the cumulative incidence function from semi-parametric and non-parametric approaches for analysing competing risk data. The investigated approaches include classical semiparametric methods such as the Fine Gray model and nonparametric methods such as random survival forests. The estimators from these approaches will be compared by simulation studies regarding bias, empirical type-I error as well as statistical power and by applying them to real world data.

**AUTHORS/INSTITUTIONS:** J. Nasejje, Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, Gauteng, SOUTH AFRICA|

**CONTROL ID:** 3390310

**TITLE:** SIMEX approach to estimation of the sparse conditional functional quantile regression with measurement error

**ABSTRACT BODY:**

**Abstract Body:** Quantile regression is a semiparametric approach used for modelling associations between variables. It is most helpful when the covariates have a complex relationship with the location, scale, and shape of the outcome distribution. Despite its robustness to distributional assumptions and outliers in the outcome, regression quantiles may be biased in the presence of measurement error in the covariates. While studies have investigated the case of scalar-valued covariates, the impact of function-valued covariates contaminated with error has not yet been examined. In this paper, we present an instrumental variable approach for consistently estimating linear quantile regression models that include a function-valued covariate measured with error. A two-stage approach to estimation is proposed. In the first stage, an instrumental variable is used to obtain a reasonable estimate of the covariance matrix for the measurement error. In the second stage, the simulation extrapolation (SIMEX) approach for measurement error correction is used to simulate additional measurement error with increasing variance which is added to the observed measure for the true function-valued covariate. The standard quantile check function is minimized after adding the simulated additional measurement error to the surrogate or observed function-valued covariate prone to error. Standard errors are estimated by means of point-wise nonparametric bootstrap. We present a simulation study to assess the robustness of the proposed estimator in the presence of measurement errors. The proposed methods are applied to the NHANES database to assess the relationship between wearable-device-based measures of physical activity on body mass index (BMI) among U.S. adults.

**AUTHORS/INSTITUTIONS:** C.D. Tekwe, Epidemiology and Biostatistics, Indiana University - Bloomington, Bloomington, Indiana, UNITED STATES|

**CONTROL ID:** 3390751

**TITLE:** Epidemiological Characteristics of Tuberculosis in Mainland China, 2005-2016: an Epidemiological Study

**ABSTRACT BODY:**

**Abstract Body:** Background The annual tuberculosis (TB) morbidity and mortality in China were higher than the world average. To better formulate prevention and control strategies, we characterized the epidemiology of TB in mainland China based on surveillance.

Methods We extracted epidemiological, clinical and laboratory data of reported TB cases during January 2005 to November 2016 and compiled geographic and demographic information. We used descriptive statistical methods to explore the spatial and temporal distribution of reported TB, and logistic regression analysis to find the risk factors of severe TB.

Results There were 10,967,235 TB cases were reported in mainland China from 2005 to November 2016 with the average incidence of 68.25 per 100 000 people. The geographical distribution of TB cases showed that cases were predominantly reported in the north, southwest and northeast of the mainland China. Residence of TB cases shifted gradually from rural to semi-urban and urban areas from 2005 to November 2016. Time from onset of illness to diagnosis with a median of 32 days, and cases from 2009 to November 2016 had a shorter time from onset to diagnosis than cases from 2005 to 2008. Male, younger people, time from onset to diagnosis more than 28 days, living in semi-urban and urban areas were high risk of severity.

Interpretation This is the largest epidemiological study of TB in mainland China, from 2005 to 2016. Future strategy of prevention and control should take full account of the concept of multi-dimensional data to explore the ecology of TB and *M. tuberculosis*.

**AUTHORS/INSTITUTIONS:** M. Liu, Capital Medical University, Beijing, CHINA|X. Guo, school of public health, Capital Medical University, Beijing, CHINA|

**CONTROL ID:** 3391040

**TITLE:** Parametric modal regression with varying precision

**ABSTRACT BODY:**

**Abstract Body:** In this paper, we propose a simple parametric modal linear regression model where the response variable is gamma distributed using a new parameterization of this distribution that is indexed by mode and precision parameters, that is, in this new regression model, the modal and precision responses are related to a linear predictor through a link function and the linear predictor involves covariates and unknown regression parameters. The main advantage of our new parameterization is the straightforward interpretation of the regression coefficients in terms of the mode of the positive response variable, as is usual in the context of generalized linear models, and direct inference in parametric mode regression based on the likelihood paradigm. Furthermore, we discuss residuals and influence diagnostic tools. A Monte Carlo experiment is conducted to evaluate the performances of these estimators in finite samples with a discussion of the results. Finally, we illustrate the usefulness of the new model by two applications, to biology and demography.

**AUTHORS/INSTITUTIONS:** M. Bourguignon, Universidade Federal do Rio Grande do Norte, Natal, RN, BRAZIL|J. Leao, Statistics, Federal University of Amazonas, Manaus, Amazonas, BRAZIL|D.I. Gallardo, Universidad de Atacama, Copiapó, CHILE|

**CONTROL ID:** 3391184

**TITLE:** Bayesian multilevel hierarchical model of nonlinear longitudinal data to understand the tumor size variability in oncology clinical development

**ABSTRACT BODY:**

**Abstract Body:** In the traditional analysis of tumor response to treatment, an overall measurement of patient tumor size is used: the sum of the longest diameters of up to 5 lesions. However, in metastatic cancer, the growth and drug-induced shrinkage of each tumor lesion may depend on the microenvironment of the hosting organ. Our aim was to develop a hierarchical model to better understand and characterize the differences in tumor size time-dynamics between lesions (within patients) and across patients. A secondary goal was to compare the anti-tumor activity of two treatment options.

The analysis dataset from the McCAVE trial consisted of individual lesion diameter (ILD) from 413 lesions in 179 metastatic colorectal cancer patients receiving modified FOLFOX-6 in addition to either vanucizumab (group 1) or bevacizumab (group 2). ILD were measured on 2D CT or MRI scans collected at baseline and every 8 weeks, until early treatment discontinuation or end-of-study.

A Bayesian 4-parameter bi-exponential model was used to describe the ILD changes over time. Inter-lesion and inter-patient variability were accounted for by adding nested random effects to the model parameters. Both treatment groups and lesion locations were evaluated as covariates to each model parameter respectively. The residual error was assumed to be additive in the log-scale and was allowed to vary across lesion locations.

The liver (80% of patients), lungs (20%) and lymph nodes (18%) were the most frequent metastatic sites. The proportion of total (non-residual) variance explained by the lesion level was ranging between 10% and 56% (depending on the model parameter).

The model goodness-of-fit was good. It revealed a faster and deeper tumor shrinkage in lung and liver compared to lesions in other locations. Also, except for lesions located in the lung, a more durable and larger effect was obtained in group 1 compared to group 2.

This modeling approach, unraveling different levels of variability, provides a better understanding of the anti-tumor drug effect in different organs and may be used to tailor treatments based on lesion location, lesion size, and early lesion response.

**AUTHORS/INSTITUTIONS:** F. Mercier, Biostatistics, F. Hoffmann-La Roche AG, Basel, SWITZERLAND|

**CONTROL ID:** 3391336

**TITLE:**

Reproducing Historical Graphics on Infectious Diseases in the Era of  
R ggplot2

**ABSTRACT BODY:**

**Abstract Body:**

Motivated by a history magazine involving statistical graphs to summarize official statistics on infectious diseases occurred in our country between the period 1923 and 1937, in this study, we would like to reproduce 10 column bar graphs in a digital platform via using R ggplot2 package. Assuming that these historical graphs were drawn by hand with a ruler, we investigated how the graphical elements of the column bars such as black and white colors, alignment of data values, legend keys, and displaying very large numbers were aesthetically designed to send the information available in the graphs to the readers.

**AUTHORS/INSTITUTIONS:** G. Inan, S. Aldag, D. Topcuoglu, Department of Mathematical Engineering, Istanbul Technical University, Istanbul, TURKEY]

**CONTROL ID:** 3391439

**TITLE:** Machine Learning Based Artificial Intelligence in Healthcare

**ABSTRACT BODY:**

**Abstract Body:** Artificial intelligence (AI) aims to mimic human cognitive functions. It is bringing a paradigm shift to healthcare, powered by increasing availability of healthcare data and rapid progress of analytics techniques. This paper gives an overview of the current status of AI applications in healthcare and discuss its future. AI can be applied to various types of healthcare data both structured and unstructured. Popular AI techniques include machine learning methods, such as the classical support vector machine, neural networks, evolutionary computing, decision trees and the modern deep learning, as well as natural language processing, fuzzy logic, rough sets, hidden Markov models, MCMC methods and related techniques. Major disease areas that use machine learning tools include non-communicable diseases such as cancer, neurology, cardiology and communicable diseases like tuberculosis, HIV and others. This paper reviews AI applications with special reference to machine learning approaches in the three major areas of early detection and diagnosis of diseases, treatment as well as outcome prediction and prognosis evaluation. Real data applications will be presented and discussed.

**AUTHORS/INSTITUTIONS:** P. Venkatesan, Bioinformatics, Sri Ramachandra University, Chennai, Tamilnadu, INDIA| P. Venkatesan, Child Trust Medical Research Foundation, CHENNAI, Tamil Nadu, INDIA|

**CONTROL ID:** 3391898

**TITLE:** Comparative analysis of false discovery rate control methods

**ABSTRACT BODY:**

**Abstract Body:** Due to the advance in technology, the type of data is getting more complicated and large-scale. To analyze such complex data, more advanced technique is required. In case of omics data from two different groups, it is crucial to find significant biomarker while controlling error rate, i.e., false discovery rate. Over the last few decades, a lot of methods that control local or global false discovery rate (FDR) have been developed, ranging from one-dimensional to k-dimensional FDR procedure. For comparison studies, we select five of them, which are unique and significant procedures: Benjamini and Hochberg (1995), Efron et al. (2001), Ploner et al. (2006), Kim et al. (2018), and Jeong et al. (Not published yet) in chronological order. The first three are one-dimensional approaches while the other two are two-dimensional ones. We compare the performance of those five methods both on simulated data and real data.

**AUTHORS/INSTITUTIONS:** S. Kim, G. Yu, J. Jeong, Statistics, Chonnam National University, Gwangju, KOREA (THE REPUBLIC OF)

**CONTROL ID:** 3392643

**TITLE:** 3D-Motion Bio-Mechanical Research of ACL Fatigue Injury Risk during Jumping and Landing Exercise

**ABSTRACT BODY:**

**Abstract Body:** What happened during 2018-2019 Warriors' Championship Series? Kevin Durant ruptured his Achilles during Game 5 and Klay Thompson tore his Anterior Cruciate Ligament (ACL) in Game 6. These two severe injuries have significantly impacted the championship result and transformed the new 2019-2020 NBA season. This paper would focus on studying the biomechanics related to Sports ACL Injury mechanism through 3D-Motion Sports Analyzer. Special designed Countermovement Force Test was conducted before and after one Fatigue Exercise. 7 Sensors were placed on the Human Body to derive the 3D-Motion Biomechanics. The fatigue factor associated with ACL Injury risk was particularly addressed through Force and Flexion profiles during both the Jumping and Landing. When body muscles getting fatigued, the body could not hold the knee steady and provide enough knee cushion during the soft toes' landing period to protect ACL in the immediate hard landing. This paper has demonstrated how to use 3D-Motion technology to help athletes monitor their vertical Jumping and Landing Bio-Mechanics, and the associated ACL Injury risk. After body being fatigued, the soft landing disappeared and the fatigued muscles could not execute Knee Flexion and provide Knee Cushion to protect ACL injury risk during the hard landing. All three Joint Flexion profiles (Hip, Knee, Ankle) have observed the fatigue impact on the landing 3D-motion and supported by biomechanics. It would be critical to detect players' fatigue status and ACL injury risk in real time during the critical games. This paper has provided a feasible methodology of measuring the players' fatigue level through a very simple vertical jump exercise and 3D-motion force and flexion measurements.

**AUTHORS/INSTITUTIONS:** M. chen, OHS, Stanford, Palo Alto, California, UNITED STATES|

**CONTROL ID:** 3392705

**TITLE:** Recursive Differencing for Estimating Semiparametric Models

**ABSTRACT BODY:**

**Abstract Body:** Controlling the bias is central to estimating semiparametric models. Many methods have been developed to control bias in estimating conditional expectations while maintaining a desirable variance order. However, these methods typically do not perform well at moderate sample sizes. Moreover, and perhaps related to their performance, non-optimal windows are selected with undersmoothing needed to ensure the appropriate bias order. In this paper, we propose a recursive differencing estimator for conditional expectations. When this method is combined with a bias control targeting the derivative of the semiparametric expectation, we are able to obtain asymptotic normality under optimal windows. As suggested by the structure of the recursion, in a wide variety of triple index designs, the proposed bias control performs much better at moderate sample sizes than regular or higher order kernels and local polynomials.

**AUTHORS/INSTITUTIONS:** C. Shen, Penn State University, Hershey, Pennsylvania, UNITED STATES|

**CONTROL ID:** 3392793

**TITLE:** Spatial-Temporal and associated predictors of unsuppressed viral load of women in a generalized hyper-endemic area in KwaZulu-Natal, South Africa.

**ABSTRACT BODY:**

**Abstract Body:** Background: Number of new HIV infections among young women in Sub-Saharan Africa remains exceptionally high, with South Africa accounting for the largest burden. To prevent onward transmission of HIV, is important to achieve UNAIDS 90-90-90 composite measure of 73% viral suppression among people living with HIV, thus potentially reducing the risk of new infections. Aim of this analyses was to spatially map the distribution of women with unsuppressed HIV viral load to help target interventions for early linkage to HIV care and scaleup treatment as prevention strategies.

Methods: Data from the HIV incidence Provincial Surveillance System (HIPSS) was analysed; two sequential cross sectional population based household surveys undertaken in Vulindlela and the Greater Edendale area of KwaZulu-Natal, South Africa. 2014 and 2015 survey enrolled 9812, 10236 participants respectively. Following informed consent, participants household global positioning system (GPS) coordinates was captured, questionnaires to obtain demographic, psycho-social behavioural information and had peripheral blood samples collected HIV related measurements. HIV viral suppression was defined as HIV VL of <400 copies per mL and mapped using R software. Survey logistic regression was used to examine predictors of unsuppressed HIV viral load.

Results: From 9812 participants enrolled in 2014 Survey, 6265 were women and 2955 (44.1%, 95% CI 42.3- 45.9) tested HIV positive, 1372 (46.4%) had a viral load of  $\geq 400$  copies per mL. Median age of HIV positive women was 32 (Interquartile range (IQR) 26-39) years, median age at sexual debut was 17 (IQR 16-19) years. Of 10236 participants enrolled in the 2015, 6341 were women and 2948 (45.0%, 95% CI (43.4-46.7) tested HIV positive, 1828 (61.9%) had a viral load of  $\geq 400$  copies per mL. Median age of HIV positive women was 32 (IQR) 26-39) years, median age at sexual debut was 18 (IQR 16-20) .

Conclusion: Substantial proportion of HIV positive women are virally unsuppressed with an urgent need for targeted interventions to link them to care, rapid scale-up of HIV treatment and prevention programs.

**AUTHORS/INSTITUTIONS:** A.O. SOOGUN, A. KHARSANY, Epidemiology, Centre for the AIDS Programme of Research in South Africa, Doris Duke Medical Research Institute, Nelson Mandela School of Medicine, University of Kwazulu-natal, Durban, Kwazulunatal, SOUTH AFRICA|A.O. SOOGUN, T. Zewotir, D. North, Statistics, University Of Kwazulunatal, Durban, Kwazulu-Natal, SOUTH AFRICA|

**CONTROL ID:** 3392808

**TITLE:** A systematic review on methods for left-censored biomarker data

**ABSTRACT BODY:**

**Abstract Body:** Classifying patients into subgroups in precision medicine strongly relies on the availability of biomarker data like gene expression profiles. Although there is a huge amount of candidate data, finding suitable profiles still is challenging due to the lack of reproducibility and statistical power. In addition, data is frequently left-censored, and it is yet unclear how best to handle data where a non-negligible proportion has values under a given detection limit. In fact, many approaches have been suggested in the literature that differ with regard to assumptions and application settings. Also, they have been investigated in different settings as defined, for instance by sample sizes, percentage of non-detects, and the underlying distribution of the data, making the practical choice difficult. In this work we therefore target this issue by summarizing published theoretical and simulation studies in which methods for the analysis of left-censored data are suggested and compared with each other. We performed a systematic review following the PRISMA statement to derive an overview of the existing methods and their applicability in different settings. The results will help to guide researchers to select the most suitable method for a specific application.

**AUTHORS/INSTITUTIONS:** D. Thiele, I.R. König, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Lübeck, GERMANY|D. Thiele, I.R. König, Airway Research Center North (ARCN), Member of the German Center for Lung Research (DZL), GERMANY|

**CONTROL ID:** 3393126

**TITLE:** Child Mortality Analysis in Ghana

**ABSTRACT BODY:**

**Abstract Body:** Infant and child mortality rates are basic indicators of a country's socioeconomic situations and quality of life. The data used for the study was from the 2014 Ghana Demographic Health Survey. The analysis of this paper is based on the women of reproductive age (15-49) years who have at least one live birth. Estimates of childhood mortality are based on information collected in the birth history section of the questionnaire administered to women. Chi-square test was employed to evaluate the association among variables associated with the mortality of the child. Multiple Logistic regression techniques as well as cross tabulation were also used to estimate the predictors. The analysis revealed that, husbands' age, region of residence, birth order, birth size, child nutritional status and accessibility to delivery centres were found to have a significant effect on child mortality. The findings further showed that parents' education, delivery at the hospital/clinic as well as type of toilet facility used have effect on a child's survival. Thus, increase in parents' education and improved health care services were seen to reduce child mortality in Ghana.

**AUTHORS/INSTITUTIONS:** A.C. MENSAH, MATHEMATICS AND STATISTICS, ACCRA TECHNICAL UNIVERSITY, Accra, GHANA|

**CONTROL ID:** 3393131

**TITLE:** Randomization and statistical methods in mouse studies

**ABSTRACT BODY:**

**Abstract Body:** Studies in humans are often justified on results of animal studies, which are frequently not designed with the same scientific rigour as are human studies; however, insufficiently planned, conducted and reported animal studies, might lead to biased results, and about 80% of molecules being shown to be effective in animal studies did not prove effective in humans. To improve the situation, statements for the preparation of protocols (DEPART) and final reports (ARRIVE) of animal studies have been published, including recommendations on randomization, allocation concealment, and blinding, which reduce bias in human studies and were shown to protect against bias and reduce effect sizes in animal studies. So far, their implementation did not yield in a substantial improvement. For instance, a review of 849 scientific papers in three journals of agricultural sciences with a focus on animal science found that completely randomized design and randomized blocks design were used in 46.5% of the animal studies, and 46.7% of the investigated studies did not use any design.

In our work, we focus on mouse studies in clinical research. Taking mice as experimental animals raises numerous statistical challenges. Born as multiples, they are kept in sex-separated cages, in groups if they are females, solitarily if they are male. Even the position of the cage in the animal station and the animal keeper can have an impact on the outcomes. These complex clustering structures should be taken into account both in randomization and statistical analyses of mouse studies. We performed a systematic literature research to summarize which randomization schemes and corresponding statistical methods are being used in mouse studies. Results show that neither mostly used randomization designs nor statistical methods are suitable for mouse studies.

**AUTHORS/INSTITUTIONS:** M.A. Vens, I.R. König, Institut für Medizinische Biometrie und Statistik, UKSH, Campus Lübeck, Universität zu Lübeck, Lübeck, GERMANY|

**CONTROL ID:** 3393158

**TITLE:** Analyzing Familial Data Using Pair-copula Models

**ABSTRACT BODY:**

**Abstract Body:** The construction of multivariate models for analyzing family data can be challenging due to the various dependencies between family members. These models typically have parameters that are subject to intricate and stringent constraints, which make it more difficult to arrive at a proper estimation of these model parameters. Recently, vines and pair-copula constructions have been utilized in order to simplify the construction of multivariate models for both continuous and discrete data. These models break down the hierarchy of dependence into vines and employ bivariate or pair-copulas as "building bricks" to generate appropriate multivariate distributions. This paper discusses pair-copula methods to model the convoluted dependence structure and proposes using a C-Vine type copula to model the dependence among family members. Analysis of actual, real-world family data is presented to illustrate our models and estimation procedures.

**AUTHORS/INSTITUTIONS:** N. Chaganty, Mathematics and Statistics, Old Dominion University, Norfolk, Virginia, UNITED STATES|Y. Deng, Mathematics and Statistics, Purdue University Fort Wayne, Fort Wayne, Indiana, UNITED STATES|

**CONTROL ID:** 3393188

**TITLE:** Randomization in Deriving Generalized Linear Mixed Models

**ABSTRACT BODY:**

**Abstract Body:** In generalized linear mixed models (GLMMs), the random effects are usually uncorrelated and assumed to be normally distributed mainly for computational simplicity. We extend the randomization approach of Brien and Bailey (2006) [Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(4), 571-609] for deriving linear models to GLMMs for the randomized complete block design (RCBD) with random block effects. The randomization of blocks and units within each block has been modelled by using the wreath product of two symmetric groups, defined by the permutation of blocks and units within each block. We also derive the variance-covariance matrices for the random effects from the randomization. In this case, the random effects are found to be correlated due to randomization. A simulation study has been conducted for estimating the model parameters using a Poisson regression mixed model.

Keywords: Randomized complete block design, Symmetric group, Wreath product, Correlated random effects.

**AUTHORS/INSTITUTIONS:** H. Grossmann, School of Mathematical Sciences, University of Magdeburg, Magdeburg, GERMANY|S. Gilmour, Mathematics, King's College London, London, UNITED KINGDOM|Z. Hossain, Statistics, University of Dhaka, Dhaka, Choose a State or Province, BANGLADESH|

**CONTROL ID:** 3393193

**TITLE:** Buttredding: A Multiple Imputation Framework to Incorporate Background Covariate Data with Sampled Outcomes for Precision Gains

**ABSTRACT BODY:**

**Abstract Body:** In medical and governmental settings, some forms of data collection are often expensive, and a small sample of the desired data is often what is only feasible. However, large administrative datasets containing background covariates on both the sampled units and a broader collection of units in the population are often readily available. This general scenario characterizes our motivating example described in Young et al. (Archives of General Psychiatry, 1998; 55: 611-617), where they wanted to determine whether a Veterans Affairs Medical Center and a Community Mental Health Center possessed the same proportions of cases of "poor-quality medication management" for patients treated for Schizophrenia. Determining whether a patient received poor-quality medication management involved clinical interviewers conducting comprehensive in-person interviews with the patients using a variety of rating scales as well as structured medical record abstractions. However, because conducting interviews is very expensive, only a small subset of the patients at both centers could be interviewed while a database of medical records was available for all patients at both centers. Such a setting can be viewed as a MCAR missing data problem where the outcome is partially observed, and the covariates are completely observed. If the degrees of association between the outcome and the background covariates are sufficiently strong, the background covariates can provide very accurate predictions of outcome values of the unsampled units, which can strengthen inferences regarding the outcome by increasing its effective sample size. Because the predictions of outcome values of the unsampled units contain error, we use multiple imputation to properly propagate this uncertainty, an approach we refer to as "buttredding," from the perspective that analyzing the observed outcomes could stand on their own as a valid analysis procedure but might be strengthened with the addition of the available background covariates on a subsuming collection of units. Using extensive simulations, we determine the magnitude of the associations between the outcome and background covariates required for buttredding to achieve meaningful precision gains, resulting in bias and interval width reductions and power gains, in the estimation of means, proportions, and their differences.

**AUTHORS/INSTITUTIONS:** J.J. Xu, T.R. Belin, Biostatistics, University of California, Los Angeles, Los Angeles, California, UNITED STATES|T.R. Belin, Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, California, UNITED STATES|

**CONTROL ID:** 3393211

**TITLE:** Productivity Assessment of Hydro-Electric Stations: An Application of Hicks-Moorsteen Total Factor Productivity Index

**ABSTRACT BODY:**

**Abstract Body:** The purpose of this paper is to study the sources of productivity change for 42 hydroelectric stations. In this paper, both the technical and technological efficiency of Indian hydroelectric stations is studied using the Hicks-Moorsteen Total Factor Productivity Index for the period of 2016-17. Out of Forty two, one station (i.e. DEPL) has achieved positive growth in all the indexes of productivity. Further, 66.67% of hydroelectric stations witnessed inefficiency in the overall productivity index. In addition, private-owned hydro stations have outperformed in all the efficiency indexes as compared to central and state-owned stations. It is also observed that 57.14% of hydroelectric stations are performing below the score of 0.5 Total Factor Productivity. This paper also contributes to the existing literature by identifying the core action area for improving the operational performance in hydroelectric stations. As a result, the policymakers will be in a better position to formulate policies or making key decisions related to hydroelectric stations.

**AUTHORS/INSTITUTIONS:** A.H. DHILLON, FINANCE, GIPCL, Kim, Gujarat, INDIA|

**CONTROL ID:** 3393238

**TITLE:** Arsenic, cadmium and chromium exposure and albuminuria: exploring causal associations under a Mendelian randomization framework

**ABSTRACT BODY:**

**Abstract Body:** Background: Albuminuria is a marker of kidney damage and microvascular disease. Exposure to metals has been associated with increased albuminuria in a number of observational studies. We evaluated the association of inorganic arsenic (iAs), cadmium (Cd) and chromium (Cr) with albuminuria in representative sample of a general population from Spain (Hortega Study), and further interrogated the causality of these associations by using a Mendelian randomization (MR) approach.

Methods: 1410 participants of the Hortega Study (Spain) had available albumin, iAs, Cd and Cr determinations in urine by ICPMS. Single nucleotide polymorphisms (SNPs) associated with genetically-elevated levels of iAs (67 SNPs), Cd (63 SNPs) and Cr (52 SNPs) were used as instrumental variables in a 2-stage least squares (2SLS) adjusted MR setting. We tested for consistency of MR results by also applying the inverse variance weighted (IVW) MR, weighted median MR and MR-Egger methods.

Results: Median levels were 3.8 mg/g for albumin, and 2.0, 0.4 and 3.5 µg/g for iAs, Cd, Cr, respectively. The observational geometric mean ratio (GMR) of albuminuria by an interquartile increase in iAs, Cd and Cr levels was 1.72 (95% confidence interval [CI]: 1.55, 1.91), 2.06 (1.87, 2.27) and 2.18 (1.99, 2.39), respectively. Using a MR approach, the causal GMR (95% CI) for iAs, Cd and Cr using the 2SLS MR method were 1.53 (1.26, 1.87), 1.83 (1.52, 2.20) and 2.15 (1.70, 2.74), respectively. These causal associations were consistent with the IVW, weighted median and MR-Egger methods.

Conclusions: Increased exposure to some heavy metals may be causally related with kidney damage at exposure levels that are relevant for the general population. The Mendelian Randomization technique is an interesting tool to disentangle causal associations where randomized controlled trials are not feasible. Replication of findings in other populations is needed to confirm these findings.

**AUTHORS/INSTITUTIONS:** M. Grau-Perez, J. Chaves, J. Redon, M. Tellez-Plaza, Institute for Biomedical Research INCLIVA, Valencia, Valencia, SPAIN|M. Grau-Perez, A. Domingo-Relloso, Autonomous University of Madrid, Madrid, Madrid, SPAIN|M. Grau-Perez, J. Bermudez, University of Valencia, Valencia, Valencia, SPAIN|A. Domingo-Relloso, A. Navas-Acien, Columbia University, New York, New York, UNITED STATES|J. Martin-Escudero, Hospital Rio Hortega, Valladolid, Valladolid, SPAIN|J. Gomez-Ariza, T. Garcia-Barrera, University of Huelva, Huelva, Huelva, SPAIN|M. Tellez-Plaza, National Center for Epidemiology, Carlos III Health Institutes, Madrid, Madrid, SPAIN|

**CONTROL ID:** 3393247

**TITLE:** Improving Predictor Generalizability Using Multiple Studies with Differing Feature Sets

**ABSTRACT BODY:**

**Abstract Body:** Typically, a set of studies or patient cohorts will exhibit heterogeneity in both the marginal distributions of the features used to train a predictor and the conditional distribution of the outcome to predict. In high-dimensional settings, we encounter the additional problem of harmonizing datasets with widely varying levels of overlap in the available features. These discrepancies can lead to poor generalization of a prediction rule learned on a single study and the discarding of potentially useful predictive features that do not fall within convenient overlaps of intersections. We describe progress made in the training of more generalizable prediction functions using multiple studies' worth of data. First, we establish preliminary theoretical guidelines and justification for Cross-Study Learning (CSL), the ensembling of predictors trained in multiple studies. We provide intuition and rules for when to merge studies together and when and how to ensemble single-study predictors. Second, we apply concepts from knowledge transfer in an effort to retain information in non-overlapping feature sets that might otherwise be discarded or ignored out of convenience. We describe conditions under which non-linear and penalized regression can be used to transfer information in the non-overlapping features using functions of the overlapping features and improve performance of a CSL predictor.

**AUTHORS/INSTITUTIONS:** P. Patil, Biostatistics, Boston University School of Public Health, Boston, Massachusetts, UNITED STATES|Y. Wu, B. Ren, G. Parmigiani, Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, UNITED STATES|Y. Wu, B. Ren, G. Parmigiani, Data Sciences, Dana-Farber Cancer Institute, Boston, Massachusetts, UNITED STATES|

**CONTROL ID:** 3393248

**TITLE:**

Apples, Pears, and Peaches – A matter of taste?

Logistic Regression, Probabilistic Index Models, and ANOVA Type Statistics  
on the Osteoarthritis Database

**ABSTRACT BODY:**

**Abstract Body:** Osteoarthritis is one of the most common chronic health conditions and a leading cause of pain and disability among adults. The purpose of the current work is to analyse the impact of age, sex, BMI, race, and contra-lateral knee pain status on the level of knee pain perception (WOMAC pain score). Due to the ordinal scale of the outcome variable WOMAC pain (0-20, 0 representing no pain, 20 the worst pain), we found three possible ways to model this which motivated a methodological discussion: We compared logistic regression models (with transformed outcome WOMAC pain yes/no), with Probabilistic Index Models (PIMs) and with ANOVA Type Statistics (ATS) for general factorial designs. In this talk, we want to discuss the advantages and disadvantages of these three models and whether it is just a matter of methodological taste.

**AUTHORS/INSTITUTIONS:** V. Racher, G. Zimmermann, A.C. Bathke, Mathematics, University of Salzburg, Salzburg, AUSTRIA|V. Racher, F. Eckstein, Imaging & Functional Musculoskeletal Research, Institute of Anatomy & Cell Biology, Paracelsus Medical Private University, Salzburg, AUSTRIA|

**CONTROL ID:** 3393259

**TITLE:** Impact of ignoring the correlation among repeated measurements when pooling studies in in meta-analyses

**ABSTRACT BODY:**

**Abstract Body:** Metanalysis and meta-regression are methods used to gain estimate efficiency with respect to those that can be achieved by an individual study. Despite their attractiveness these techniques poses several challenges. One of those is represented by the correlation that might affect the observations collected within an individual study. This is the case for instance of baseline and post-treatment measurements taken on the same group of individuals in a parallel trial or in any crossover trial where individuals are allocated to two different treatment groups and outcome measurements are taken twice on them. In these situations the effect measure on an outcome (e.g. the serum triglycerides difference due to a different intake of a macronutrient) is computed as the difference of the outcome mean change from baseline between the treatment and the control group (parallel studies) or the difference of the outcome mean at the end of each treatment type (cross-over studies). In order to estimate the standard error of the mean effect with no bias, an estimate of the correlation coefficient is required. A positive correlation (the most probable to observe) leads to a standard error that is lower than the one estimated under the assumption of independence. When doing the pooling, this turns into a larger weight assigned to the studies with correlated observations both under fixed and random effects meta-analysis. Unfortunately, because of the poor reporting, the information related to the correlation coefficient in repeated observations in a study is rarely made available in a paper. As a consequence independence among observations is assumed even when not appropriate or correlation is simply ignored in several meta-analysis and meta-regression. This paper discusses the possible impact of overlooking the correlation across observations in metanalysis. Using a probability distribution to express the uncertainty on the correlation parameters in parallel and cross-over studies, random values are simulated and used to compute the outcome estimate precision and the weight of true randomised controlled trials published in the scientific literature. The meta-analytic pooled estimate, its precision and power are computed both under the assumption of independence and taking the correlation into consideration. The results and their differences are discussed.

**AUTHORS/INSTITUTIONS:** L. MARTINO, V. Ercolano, Dept. of Risk Assessment and Scientific Assistance, European Food Safety Authority, Parma, ITALY]

**CONTROL ID:** 3393266

**TITLE:** Modeling and simulation of the insecticide potential analysis with a model of mixtures of bioactive constituents present in essential oils

**ABSTRACT BODY:**

**Abstract Body:** Cereals play a crucial role in the process of ensuring food security. However, these must face storage pests that cause significant losses, which are traditionally controlled with pesticides and cause negative impacts on human health and the environment, in addition to the generation of resistance to this type of controls by the insect. Therefore, the objective of this study was to take advantage of compounds of natural origin, which are part of the composition of certain essential essential oils on these pests. Initially, the individual activity of these constituents was evaluated through a probit regression model, to determine a set of parameters that allow clustering of the compounds. Based on the previous results, the experimental mix quantity design and statistical response surface modeling was used, compared to the simulation of the combination index to determine the binary mixtures that show synergism characteristics in the fumigant activity. We suggest that some combinations of compounds could contribute to the management of these pests and show that the experimental design of the amount of mixture and the index of combination is a useful tool for this type of experiments

**AUTHORS/INSTITUTIONS:** E. Herrera, mathematics, Pontificia Universidad Javeriana, Bogotá, Cundinamarca, COLOMBIA|L.J. Nagles, L. Perdomo, O. Patiño, Quimica, Universidad Nacional de Colombia, Bogota, Cundinamarca, COLOMBIA|J. Prieto, Quimica, Pontificia Universidad Javeriana, Bogota, COLOMBIA|

**CONTROL ID:** 3393294

**TITLE:** Variable Selection in Nonparametric Additive Quantile Regression for Genomic data with Prior Information

**ABSTRACT BODY:**

**Abstract Body:** A priori information, such as biological pathways, is a useful supplement in identifying risk factors of a trait using genomic data. However, the commonly used methods to incorporate prior information provide a model for the mean function of the outcome and rely on unmet assumptions. To address these concerns, we propose a method for variable selection in nonparametric additive quantile regression with network regularization to incorporate the information encoded by known networks. To account for nonlinear associations, we approximate the unknown additive functional effect of each predictor with the expansion of a B-spline basis. We implement the group Lasso penalty to obtain a sparse model. We define the network-constrained penalty by the total  $\ell_1$  norm of the difference between the effect functions of any two predictors that are linked in the known network. We further propose an efficient computation procedure to solve the optimization problem that arises in our model. Simulation studies show that our proposed method performs well in identifying more truly associated variables/genes and less falsely associated variables/genes than alternative approaches. We apply the proposed method to analyze the microarray gene-expression dataset in the Framingham Heart Study and identify several body mass index associated genes. In conclusion, our proposed approach efficiently identifies the outcome-associated variables in a nonparametric additive quantile regression framework by leveraging known network information.

**AUTHORS/INSTITUTIONS:** P. Wu, J. Dupuis, C. Liu, Boston University, Boston, Massachusetts, UNITED STATES|

**CONTROL ID:** 3393347

**TITLE:** Analysing causal effect of London cycle superhighways on traffic congestion and air pollution

**ABSTRACT BODY:**

**Abstract Body:** Transport operators have a range of intervention options available to improve or enhance their networks. But often such interventions are made in the absence of sound evidence on what outcomes may result. Cycling superhighways were promoted as a sustainable and healthy travel mode which aims to cut traffic congestion and air pollution. The estimation of the impacts of the cycle superhighways on congestion and air pollution is complicated due to the non-random assignment of such intervention over the transport network. In this paper, we analyse the causal effect of cycle superhighways utilising pre-intervention and post-intervention information on traffic and road characteristics along with socio-economic factors. We propose a modeling framework based on the propensity score and outcome regression model. The method is also extended to doubly robust set-up. Simulation results show the superiority of the performance of the proposed method over existing competitors. The method is applied to analyse a real dataset on the London transport network. The methodology proposed can assist in effective decision making to improve network performance.

**AUTHORS/INSTITUTIONS:** P. Bhuyan, Mathematics, Imperial College London, London, UNITED KINGDOM|

**CONTROL ID:** 3393391

**TITLE:** On the estimation of population size under dependent dual-record system: a new integrated likelihood approach

**ABSTRACT BODY:**

**Abstract Body:** Motivated by various applications in the domains of epidemiology, population studies, criminology, etc. the problem of estimating size of a homogeneous human population based on two-sample capture-recapture experiment is considered in this article. The Lincoln-Petersen estimate assuming causal independence between the sources is widely used though relevance of this assumption is unanimously criticized in most of the applications. We compute the accuracy and efficiency of this estimator under plausible existence of behavioral dependence among the sources or samples. Efficient estimation of the population size from this dependent dual-record system (DRS) remains a statistical challenge in the capture-recapture type experiment. In this context, Time-Behavioral Response Variation model is most relevant. Owing to the non-identifiability of the suitable time-behavioural response variation model under DRS, few methods are developed in the Bayesian paradigm based on informative priors. This article theoretically investigates that basic integrated likelihood fails to make inference from this model. Our contribution in this article is to develop a new integrated likelihood function from this model motivated by a novel approach developed by Severini (2007). A suitable weight function on the nuisance parameter is derived with the knowledge of the direction of behavioural dependency. A pseudo-likelihood function is constructed so that the resulting estimator possess some desirable properties including negligible prior (or weight) sensitiveness. Extensive simulations show the superior performance of our proposed method to that of the existing Bayesian methods. Moreover, the proposed estimator is easy to implement from the computational perspective. Finally two real life data with different characteristics are analyzed as practical illustrations of the proposed method.

**AUTHORS/INSTITUTIONS:** K. Chatterjee, Department of Statistics, Bidhannagar College, Kolkata, India, Kolkata, INDIA|D. Mukherjee, Sampling and Official Statistics Unit, Indian Statistical Institute, Kolkata, Kolkata, INDIA|

**CONTROL ID:** 3403519

**TITLE:** Added value of biomarkers and polygenic risk scores as risk factors for coronary artery disease

**ABSTRACT BODY:**

**Abstract Body:** Traditional risk factors such as BMI and smoking are assessed in the process of coronary artery disease (CAD) diagnosis for individuals. At a population level, stratifying by CAD risk could allow for more targeted approaches to prevention, diagnosis and treatment. Biomarkers, such as cholesterol, can also be used as risk factors. However, these are often only detectable once the disease pathology has begun, so their value as a predictive measure for CAD is likely to be limited by proximity to age of onset. Polygenic risk scores (PRS), which are a single measure of an individual's genetic likelihood of developing a disease, remain constant across the lifespan. Thus, they could potentially help identify the groups at highest risk of CAD and allow for earlier management and intervention. To investigate the value biomarker and genetic data could add to risk models in addition to traditional risk factors, we used the UK Biobank to build and compare Cox proportional-hazards regression models including these data. The area under the curve (AUC) given by our traditional risk factor model (0.734) was modestly improved by the inclusion of biomarkers (AUC 0.738) and PRS (AUC 0.748), and our full model including all the potential risk factors achieved an AUC of 0.750. This suggests that biomarker and genetic data hold value alongside traditional risk factors for CAD and could improve stratification for more targeted interventions. The inclusion of these data in risk models could potentially allow for better prevention and diagnosis of CAD.

**AUTHORS/INSTITUTIONS:** N. Sharapova, J.M. Maxwell, K. Glanville, S. Hagenaaars, C.M. Lewis, Social, Genetic and Developmental Psychiatry Centre, King's College London, London, UNITED KINGDOM|R. Russell, Global Research and Data Analytics, RGA Reinsurance Company, London, UNITED KINGDOM|Z. Ibrahim, Department of Biostatistics & Health Informatics, King's College London, London, UNITED KINGDOM|